

Assessment Task Force Report

May 31, 2006

**Created by the State of Delaware
House Joint Resolution No. 4
with Senate Amendment No. 2**

State of Delaware
House Joint Resolution No. 4 with Senate Amendment No. 2
Assessment Task Force Report

Executive Summary

During the 143rd General Assembly, Representative Nancy Wagner and Senator David Sokola were the prime sponsors of House Joint Resolution 4 establishing a task force that would review the best practices of educational assessments for the purpose of providing recommendations for the improvement and/or changes to the Delaware Student Testing Program (DSTP). The task force met eleven times beginning with the first meeting on December 15, 2005 and ending on May 25, 2006.

During the process, the task force discussed the current assessment program and considered what the assessment program might look like in the future. Various innovative approaches were brainstormed that included existing vendor-based assessments as well as customized comprehensive assessment systems. National experts were brought in at the request of the task force members to discuss core assessment requirements, such as the importance of aligning the assessment to Delaware's content standards; information on other state assessments; computer-based assessments; and test validity.

The task force decided that an assessment system will be recommended to the Governor and General Assembly that consists of two major components: locally-administered formative assessments and state-administered annual summative assessments. Reading, writing, and mathematics will be particularly emphasized in grades two through eight in both a formative and summative manner. Summative assessments will be given in grades three, five, and eight in science and social studies annually. Formative assessments in science and social studies will also be given periodically. As for high school, end-of-course assessments are the preferred choice for the summative piece and again, formative assessments to be given systematically in reading, writing, mathematics, science and social studies. The summative assessments in all content areas will consist of both multiple choice and constructed response test questions. The constructed response test items will be given in the Spring and the multiple choice test items will be given as close to the end of the school year as possible in May.

Scoring and reporting are other important components. Scoring could be done through technological means; if possible, Delaware teachers could be used as scorers for constructed response questions. Results of the assessments are to be reported as quickly as possible. All of these areas should work to inform instruction and focus on professional development. Some components, but not necessarily all components, could be used to determine accountability status.

The full report from the Assessment Task Force follows, including a statement of purpose for the assessment system.

Background

During the 143rd General Assembly, Representative Nancy Wagner and Senator David Sokola were the prime sponsors of House Joint Resolution 4 establishing a task force that would review the best practices of educational assessments for the purpose of providing recommendations for the improvement and/or changes to the Delaware Student Testing Program (DSTP).

The task force membership consisted of the following:

1. Representative Nancy Wagner, Chairperson of the House Education Committee, Delaware House of Representatives;
2. Senator David Sokola, Chairperson of the Senate Education Committee, Delaware State Senate;
3. Vicky Cairns, Delaware State Education Association (DSEA);
4. Nancy Doorey, Delaware School Boards Association (DSBA);
5. Dave Sechler, Delaware Association of School Administrators (DASA);
6. Bruce Harter, Chief School Officers Association (CSOA);
7. Jean Allen, State Board of Education;
8. Cindy DiPinto, Delaware Business Roundtable;
9. Janine Sorbello, Representative, Delaware State Chamber of Commerce;
10. Emily Falcon, State Budget Office;
11. Martha Manning, Delaware Charter Schools Network;
12. Dorothy Shelton, District DSTP Test Coordinator;
13. Yvonne Johnson, Parent;
14. Noel Rodriguez, Parent;
15. Edie Corbin, Metropolitan Wilmington Urban League;
16. Valerie Woodruff, Secretary of Education
17. Jim Wolfe, President/CEO of the Delaware Chamber of Commerce was named to the committee; however, Janine Sorbello served as his proxy;
18. Bob Andrzejewski, Chief School Officers Association, attended the last two meetings of the task force

The task force met eleven times beginning with the first meeting on December 15, 2005 and ending on May 25, 2006.

During the process, the task force discussed the current assessment program and considered what the assessment program might look like in the future. Various innovative approaches were brainstormed that included existing vendor-based assessments as well as customized comprehensive assessment systems. National experts were brought in at the request of the task force members to discuss core assessment requirements, such as the importance of aligning the assessment to Delaware's content standards; information on other state assessments; computer-based assessments; and test validity.

For purposes of this document, these four terms have the following meaning:

Formative assessment: includes all activities that are used to provide information to teachers on the progress a student is making in a particular subject or course to help the teacher direct the learning activities toward mastery of the skills and knowledge intended to be learned in that subject at that time or in that course. The main purpose is to help the individual student master the content standards through course materials. The formative assessment will also be used to show growth in student achievement throughout the school year.

Summative assessment: includes any activities that are given at the conclusion of a course or an instructional period (such as Spring of Grade 5) to determine the success of the instruction up to that point in time. The summative assessment will also be used for student, school, district, and state accountability.

National norms: are established by testing a national sample representing a wide and diverse cross-section of students -- a "norm group." Students, schools, districts, and even states are compared or rank-ordered in relation to the norm group. The purpose of a norm-referenced assessment is usually to sort students and not to measure achievement towards some criterion of performance.

Infrastructure: for the purposes of this report, refers to the technology and personnel necessary for implementation of the assessment system.

National Experts and Discussion Topics

The experts and the topic of discussion with the task force included:

National Expert	Credentials	Topic
Brian Gong	Executive Director, National Center for the Improvement of Educational Assessment	Federal Requirements and The Future of Assessment Programs
Ellen Forte	Independent Consultant, former President, edCounts, LLC, former Director of Student Assessment, Baltimore City Public Schools	What Other States are Doing/Planning How Delaware Can Operationalize the Purpose Statement
Ron Hambleton	Professor, Chairperson of the Research and Evaluation Methods Program, Co-Director of the Center for Educational Assessment, University of Massachusetts at Amherst	Computer Based Assessments
Lou Fabrizio and staff	Director, Accountability Services Division North Carolina Department of Public Instruction	North Carolina's Assessment System with a focus on End of Course Assessments
Steve Dunbar	Professor, College of Education, University of Iowa; Chairperson of the Delaware Technical Advisory Committee (TAC)	The Importance of Validity and Reliability in Assessments
Phoebe Winter	Independent Consultant, frequently works with Council of Chief State School Officers (CCSSO)	Alignment Study for Formative Assessments

Statement of Purpose

It was very important from the discussions and information from the experts that the task force create a purpose statement. This statement would articulate why and how the assessment system should be used.

The following is the statement of purpose as decided by the task force:

The primary purpose of the next generation of the Delaware Student Testing Program should be the improvement of student learning.

The State of Delaware's student assessment program should be a system of assessments that fairly and accurately measure student achievement against state content standards.

The system should incorporate components that:

- Measure progress of student achievement toward the standards at regular intervals.
- Provide timely and definitive information to educators, especially teachers, to inform instruction.
- Provide timely information to parents and families regarding achievement of their children so that they are empowered to assist the children's learning.
- Provide information that will assist districts and charter schools and the state to measure the impact of instructional programs and practices.
- Allow for the establishment of individual student goals and accurately measure each child's growth over time based on the standards.
- Provide any required elements for federal and state accountability.

The system should measure individual student progress and have the necessary components to provide data to measure student and system performance. These data should also be used, with other measures, to inform evaluation of educator performance. Such information should be used to guide decisions regarding individual professional improvement as well as system wide improvement. It should provide the public with a fair evaluation about the growth in student achievement occurring in each school and district.

One assessment given at a single point in time during the school year cannot meet the goals outlined above. One component of the system should include assessments that provide periodic and timely feedback regarding the yearlong learning process, which is critical to the success of our children and our system. There must also be a component

that serves as an annual measure of system performance and that contains a measure of performance based on national norms.

The system shall be designed and implemented mindful of cost efficiencies and attentive to the amount of testing time necessary to obtain valid and reliable data. The Department of Education shall obtain feedback annually on the assessment system, provide such feedback to a broad based group of stakeholders to include district level educators, teachers, and families for the purpose of seeking the group's recommendations, summarize changes and enhancements to be made, and communicate this to various constituency groups.

Accountability – Federal and State

The task force was cognizant of the federal requirements as well as our current state student accountability requirements. As background, the state is required to be in compliance with the reauthorization of the federal Elementary and Secondary Education Act of 1965, commonly referred to the No Child Left Behind Act (NCLB) of 2001. This Act requires that each state must have a valid and reliable assessment system that measures individual student progress towards the state standards in grades 3 through 8 and one grade in high school (not grade 9) in reading and mathematics. Assessing students in science at least at the elementary, middle and high school levels must to be included by 2007--08.

In addition, all students are required to take the same assessments at the grade in which they are enrolled with the exception of a 1% cap of the students with the most severe cognitive disabilities. Another 2% of the students with disabilities may take an alternate assessment. At this time, the US Department of Education has not issued regulations governing this assessment; therefore, Delaware will not be able to develop such an assessment until such time as appropriate guidance is provided by USDOE. NCLB requires that states provide for "reasonable adaptations and accommodations for students with disabilities" and for "the inclusion of limited English proficient students who shall be assessed in a valid and reliable manner and provided reasonable accommodations." The accountability system is based on performance in the following subgroups: All students, each racial/ethnic group, students with disabilities, low income, and students who are English language learners. The overarching goal is that all students will be proficient in reading and mathematics by the end of the 2013-2014 school year.

Each state is required to build a valid and reliable accountability system and to go through a peer review process to determine if the state's standards and assessments meet criteria as established by the USDOE. Delaware's current standards and assessments were approved by the federal USDOE in March 2006.

The task force was also aware of the student accountability provisions such as retention, summer school and the awarding of a diploma. The task force did not focus on these requirements as limitations. The intent of the task force was to provide a

system that would inform students, teachers, and families in the individual student learning process. The system would be designed to assist students, teachers, and administrators with curricular and instructional decision making.

The “Next” Assessment System

The task force decided an assessment system will be recommended to the Governor and General Assembly that consists of two major components: locally-administered formative assessments and state-administered annual summative assessments. Reading, writing, and mathematics will be particularly emphasized in grades two through eight in both a formative and summative manner. Summative assessments will be given in grades three, five, and eight in science and social studies. Formative assessments in these content areas will also be given periodically and will be embedded within curriculum units. As for high school, end-of-course assessments are the preferred choice for the summative piece and again, formative assessments to be given systematically.

Scoring and reporting are other important components. Scoring could be done through technological means; however if possible, Delaware teachers could be used as scorers for constructed response questions. Results of the assessments are to be reported as quickly as possible. All of these areas should work to inform instruction and to focus on professional development. Some components, but not necessarily all components, could be used to determine accountability status.

Description of Proposed Delaware Assessment System

To meet the purposes identified by the task force, the new statewide assessment system would comprise two major components:

- 1) For the purpose of informing instruction, the state would provide an infrastructure* for locally-administered formative assessments.
- 2) For accountability purposes, the state would administer annual summative assessments in grades three through eight and end-of-course assessments at the high school level.

This new assessment system would be customized for Delaware so that our stakeholders retain control over the content, structure, administration, scoring, and reporting of our tests. Our assessment process would help close the gap between our tests and our classroom practices by engaging Delaware educators in the design and construction of tests that are truly aligned with our standards and their curricula. Our new system would represent a comprehensive, integrated approach to instructionally-supportive assessment that brings together curriculum, instruction, and measurement of

* Infrastructure refers to the technology and personnel necessary to implement the assessment.

learning that enables continuous improvement in the implementation of standards in classrooms across the state.

Together, the formative and summative accountability assessments would balance high stakes performance expectations with the information necessary to support teaching and learning. By designing these two components as parts of a single system, we would provide more consistency across testing circumstances. In addition, students, educators, and families would receive coherent information about student performance throughout the year rather than statewide assessment reports that seem to be unrelated to reports of student progress during the school year. Although the content of formative and summative accountability reports would differ in detail, the structure, language, and format of the reports would be consistent.

The elements that would be common across these components are described first below. Each component is then described in further detail in the sections that follow.

Common Elements

While the two components of the new statewide approach to assessment would be designed to serve different purposes, they would share several elements:

- **Alignment to Delaware Content Standards in English Language Arts (Reading and Writing), Mathematics, Science and Social Studies**
We have chosen not to narrow our educational focus in Delaware. Including assessments in these four core content areas supports the high value we place on each of these areas. Alignment with standards ensures that our assessments reflect the full breadth, depth, and patterns of emphases of our expectations.
- **Use of multiple-choice and constructed-response types of questions**
We value students' higher-order thinking and communication skills. To enhance alignment with our grade-level expectations and assess these types of skills appropriately, our assessments will include constructed-response questions as well as multiple-choice questions.
- **Use of computer-based technology for test administration**
We encourage the use of assessment data for educational decision-making and must, therefore, return assessment information to educators, parents, and students as quickly as possible after testing. Using computer-based technologies for administering our assessments will reduce the time necessary for scoring students' responses and returning results to schools. In addition, testing by using technology will better match the growing use of computers in our classrooms and the educational and professional contexts our students will enter upon graduation from high school.
- **Involvement of Delaware educators in the assessment process**
We value the expertise of Delaware educators and view all assessment as serving instruction directly or, through accountability mechanisms, indirectly. We

understand that involving educators in the assessment process for our statewide tests draws upon our teachers' content and instructional knowledge and helps to broaden their assessment literacy. While the task force believes educators' scoring of students' responses is valuable, it is not a mandatory component of the system. The task force would like to explore continuing the practice of involving our teachers in item development and benchmarking, and considerations should be made for involving teachers in scoring. Serious concerns were raised by task force members about the advisability of teacher involvement in scoring that would encroach upon instructional and professional development activities.

- **Reporting designed to support interpretation and use by educators, families, and students**

We believe that the primary purpose of testing is to provide educators, families, and students with meaningful information that they can use to understand and support academic achievement. Therefore, we must communicate test results clearly, accurately, and as soon as possible after each test administration.

Formative Assessment

In Delaware, as in most other states, school districts and charter schools are responsible for the instructional, curricular, and professional development programs that directly impact teaching and student learning. While many large districts are able to direct resources into buying or developing academic assessments that support their work, some smaller districts and charter schools are not able to capitalize on assessment options that require extensive investments of time and money. In addition, assessment options that are commercially available to districts are generally not well-aligned with Delaware's grade-level expectations; thus, the results of these assessments may not provide information that is directly relevant to educators' curricula and decision-making needs. The formative component of our system is intended to address these issues by offering equity of access to aligned assessment options for all districts and charter schools across the state.

Our approach would not dictate the tests or test questions that districts and charter schools administer to their students. Rather, state-supported formative assessment would encompass the following:

- Formative assessments that would be built and would include items purchased from commercial sources, developed in collaboration with other states and/or items developed by teachers from across Delaware. Districts and charter schools retain control over when and how to use this component. The state will ensure availability for all districts and charter schools;
- Formative assessments that measure growth in individual student achievement from fall to spring and year to year while measuring their performance along the continuum of our grade level expectations.

- Computer-based as well as computer-adaptive administration of these assessments to provide immediate feedback in the detail necessary to support specific instructional actions for individual students and to show individual student growth;
- Availability of these assessments in paper and pencil version for students who need such accommodations;
- Assessments for Science and Social Studies embedded in the Science and Social Studies Units – Statewide Recommended Curriculum;
- Immediate computer-scoring and feedback from multiple-choice items;
- A variety of local options can be used for scoring constructed response items that are embedded in the curriculum units, from individual teachers scoring their own students' responses to the creation of school-wide or district-wide grade level teams that build consensus on performance criteria;
- Access to professional development programs for assessment literacy designed and delivered collaboratively by the districts and charter schools and the state to support the creation of high quality questions for both classroom and statewide use, the construction of tests, and the appropriate interpretation and use of results.

Statewide Summative Assessments

The purpose of Delaware's summative assessments system is to ensure that programs are effectively supporting student achievement and that resources are directed to schools and districts or charter schools appropriately. The results of our annual statewide summative assessments are critical indicators in the state and federal accountability system as they provide a single, common yardstick for measuring our grade-level expectations for all Delaware students.

The statewide summative assessments would encompass the following:

- Assessment of Reading, Writing, and Mathematics grade-level expectations at grade-levels two through eight;
- Assessment of Science and Social Studies end-of-cluster expectations of the content standards at the end of grades three, five, and eight;
- Assessment of achievement expectations in Reading, Writing, Mathematics, Science, and Social Studies at the end of specific courses at the high school level (specific courses to be determined later);
- Administration of multiple-choice questions as close to the end of the school year as possible while allowing adequate time for re-takes and the ability to produce reliable, valid scores for every student and school and make subsequent accountability decisions;

- Administration of constructed-response questions in the spring of each year to allow necessary time for scoring these assessments while still returning reports back to educators, families, and students by the end of the school year;
- Administration of assessments via computer, but with paper-and-pencil versions available for students who need such accommodations and for some constructed-response questions;
- Scoring of constructed-response questions by involvement of a combination of Delaware educators and external scorers to balance cost and time efficiencies with the benefits of local insight and involvement with the high stakes scoring processes;
- Annual release of some multiple-choice and some constructed-response questions along with each student's responses for professional development purposes and to enhance families' and students' understanding of test content and format; and
- Administration of re-takes of the assessment in the summer, as appropriate.



Together, these components would provide the multiple types of information we need to make important decisions about student achievement and how to support it. We recognize that tests themselves cannot improve achievement. Rather, we effect change only when we place primary focus on our standards and grade-level expectations, understand how to build aligned items and tests, and learn how to score and appropriately use information from these tests.

Closing Comments

The task force recommends that the next step should be the development of a Request for Proposal (RFP) for the next generation of the Delaware Student Testing Program that encompasses the components as described in this final report. The task force also would like to express to the Governor, the General Assembly and citizens of Delaware that the new system may take additional time and funds for full implementation. It is the intent of this task force to provide the framework for an assessment system that will ultimately provide educators, families and students the information needed to make sound instructional decisions and that provides information on where students stand against the state content standards over time.

Next-Generation State Assessment System Update

G/B	Formative assessments that measure growth in student achievement: 45 – 60 minutes	EOG	Summative End of Grade/End of Course assessment: computer-based, 45 – 90 minutes
	Constructed response items		Writing assessment

Grade	Sept	Oct	Nov	Dec/Jan	Feb	Mar	Apr	May/June
2 – 8	G/B	Available on demand up to three times per year						EOG G/B
				G/B			Up to mid-April	
9-12	G/B				EOC		Up to mid-April	EOC
	Optional, available in reading and math, up to 3 times per year							

Grades 2-8: Formative assessment in reading and math given on demand at beginning of school year and up to two more times.
 Grades 3, 5 and 8: Science and Social studies EOG tests given in final month of school or in fall of grades 4 and 6, final months of school in grade 8
 Constructed response items: short response items in each content area, plus writing assessment, given in spring

Grades 9 – 12 Formative assessments available for use in reading and math
 At least 5 EOCs required before graduation: courses TBD, but at least 1 writing, 1 math, 1 reading, 1 science, 1 social studies
 Taken as close to end of year as possible, with 2nd version available in final weeks for those who do not pass on first try
 Results available as quickly as possible
 Constructed response items: short response items per content area, plus writing assessment, given in spring

NOTE: EOC to be given at grade in which student takes the course (e.g. Algebra I in grade 8)

APPENDIX A

Meeting Notes

HJR 4 with SA 2 – Assessment Task Force

Meeting Notes

May 25, 2006

Location: Library Conference Room, Delaware Department of Education

Time: 1:30 p.m.

Attendees: Bob Andrzejewski, Jean Allen, Vicky Cairns, Edie Corbin, Cindy DiPinto, Nancy Doorey, Bruce Harter, Yvonne Johnson, Martha Manning, Nicole Quinn, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Janine Sorbello, Nancy Wagner, Valerie Woodruff,

Public: Jon Twing and Brian Farmer from Pearson Educational Measurement, Amy Drevna from Harcourt, and Cecilia Le from the News Journal

Secretary Woodruff called the meeting to order. She wanted to make sure that everyone had a hard copy of the ten page “HJR 4: Assessment Task Force Final Report” which was to be reviewed in the meeting. The goal of the meeting was to finalize the report and send it to the General Assembly and the Governor next week. A re-write was sent by Nancy Doorey and Bruce Harter and the group did a side-by-side of the two documents using the “Quick Reference: Differences between the Option A Assessment Task Force Report and the Alternate “Option B” Report”.

Point number one on the Quick Reference Document discussed offering the test up to three times per year. This raised concerns for Mrs. Woodruff regarding the added security that would be needed to offer more than one high stakes summative assessment for accountability per year. This would add not only security, but be an issue of testing time and additional costs for administration and development. Nancy Wagner wanted to make it clear that she could not support anything that raises cost or spends more time testing in schools. Nancy Doorey pointed out that the decision to test more than once should be made at the school level, where it would be available but not required. However, Mrs. Woodruff was concerned about the opportunity problem or equity among districts that might occur. We cannot allow one school to offer it three times and another school to only offer it once. It must be offered equally across the board or parents will be very upset. It cannot be a loose system. In addition, if we offer the test multiple times then we have multiple security issues which will drive up cost.

Mrs. Cairns asked about the provisions for the ability to retake the summative assessment. Mrs. Woodruff responded by stating that the multiple choice portion will be given as close to the end of the school year as possible, such as mid-May. That would allow enough time for a retake for those that need it. These scores would be used for AYP. Also, students could take the test at the end of summer school. However, this test would be a different summative test and, as of now, those scores cannot be used for accountability. She again mentioned that offering the summative test three times a school year would be a major issue.

A clarity question was asked regarding what circumstance the student would not get the same summative exam that is given during the standard exam time. Mrs. Woodruff stated that the summer school summative exam would be the only different exam offered. If a student is sick during the standard exam time, he or she would take the retest which is the same exam as the one given prior.

Mrs. Wagner and Mrs. Woodruff agreed that offering the summative exam three times a year would probably not be welcome by the General Assembly.

Cindy DiPinto added that she likes the proposal (referred to as Option A during the meeting) on the table. It keeps all stakeholders vested till the end of the school year. The group agreed that the summative exam would remain as proposed in the Option A proposal.

The group moved to point number two on the Quick Reference Document. Mrs. Doorey pointed out that if we want to measure growth it must be done as a fall-to-spring adaptive growth measure. With the current DSTP assessment, we cannot get a handle on “growth” other than the student’s performance level. Mrs. Doorey elaborated by stating that the second test would be an adaptive growth measure that would not be used for accountability but would be used for students to see how they are performing. She stated that the student would take the adaptive test in September to get an accurate assessment of performance and in the spring the student would take the summative assessment as well as an adaptive assessment in one sitting to get the growth information. The transition between the two tests would be seamless for the student. The student would then take the adaptive test in the fall to see what happened to learning over the summer. Mr. Sechler mentioned that for DPAS purposes, you could look at results through a school year to see growth for the set of students the teacher has. Mrs. Cairns added that there is a need for some sort of adaptive assessment through the school year that will inform instruction. Mrs. Woodruff stated that Option A does have the two-part summative and formative exams that require more testing however it is a looser format than what Option B is proposing. She stated that calculation of growth would have to be summative. Mr. Sechler did not understand what the difference was between what Option A proposed and what Option B was proposing. He stated that the DSTP does calculate student growth however school growth is more difficult. He does not understand the reference of school growth in point two.

There was confusion in the group about the wording used in point two on the Quick Reference document regarding school growth. Mrs. Doorey agreed that “student or school” should be removed from the Quick Reference document and it should just refer to growth. Also, the term “calculation” was an issue.

Mrs. Shelton pointed out that the group should keep formative and summative assessments separate for now. Mrs. DiPinto stated that she does not see anything that ties formative to instructional practice. Mrs. Johnson also mentioned that the report from the assessment should give specific details where the student is weak. Mr. Sechler stated that we have different tests that serve different purposes. The DSTP says if a student is meeting the standards. The formative tests will find the students’ break point. That will inform instruction. He thought that the Option A proposal had these two components and does not think changes should be made.

Mrs. Doorey stated that not all formative tests give growth information. We need an adaptive growth measure. She stated that the Option A proposal does not say this and it needs to be more specific.

Mrs. Cairns clarified that the group agreed to the following: There should be adaptive tests to inform instruction two or three times per year then there is a summative test once a year.

Mr. Sechler asked about the timeline that these tests will be offered. Mrs. Woodruff stated that the RFP will need to say what we envision and the vendors will tell us how it will work practically. She also stated that another piece of the RFP would need to include the national norm comparison component. Right now, we might be asking for things that do not exist.

The group went to point three in the Quick Reference document that dealt with end-of-course exams. Mrs. Woodruff stated that this is “to be determined” because the graduation task force has not yet completed its work. Mrs. Allen mentioned that the issue with end-of-course brought up in the graduation task force is that of courses taught in the traditional or integrated approach. The graduation task force feels that the district should be able to choose which style it would want however this would make it more difficult regarding end-of-course exams. Mrs. Woodruff replied that, in this case, two end-of-course exams would then have to be offered. For example, one would be offered for Algebra I and another offered for Integrated II. However, the challenge would be that the two different tests would have to measure the same skills and cover the same content areas/grade-level expectations (GLEs).

Mrs. Doorey raised some concerns that she has heard from other stakeholders with the integrated approach. She said that some students were getting rejected from college because the college did not understand what was being taught in the integrated classes. She also said that for those students moving into Delaware, traditional courses are a better option. Mrs. Woodruff replied that students are moving into Delaware everyday and we have to deal with them in all kinds of ways. As for college acceptance, this is a simple fix by including a description on the transcript regarding the course. She mentioned that the University of Delaware accepts integrated coursework and they are very competitive with admissions. As long as the two end-of-course exams measure the same skills and are used for accountability it will work. However, we cannot substitute other exams such as the AP exams in their place.

Mrs. Cairns asked if the end-of-course exams will be part of the final grade, the graduation requirement, and/or accountability. Mrs. Woodruff replied that for the purposes of this task force, they will be used for accountability. However, it is important that the student be vested in the exams so it could count as part of the students’ grade as well.

Mrs. DiPinto was concerned about the writing portion and the group stated that writing is a part of the process throughout. Also mentioned and clarified is that end-of-course does in fact mean the exam is given at the end of the course not at the end of the year.

The proposal should broadly state that “the task force wants to foster and support end-of-course tests.”

The group moved to point four of the Quick Reference document that dealt with cost and timing. Mrs. Woodruff made clear that the task force is not proposing a cheaper system and it needs to be clear that the system will be efficient and effective with cost. Regarding time, Mr. Sechler stated that this is an RFP issue, not a task force issue. He stated that reducing the testing time is a goal however an appropriate amount of time is needed to accomplish the end goal.

Mr. Harter asked if special education and accommodations could be added to the report. He also asked that it be clear in the final report that these changes will need to be phased in over time.

The group moved to point five of the Quick Reference document that dealt with a “customer-developed set of assessments”. Mrs. Woodruff agreed that “customized” might be better language. There is no intent that all the assessments be custom-developed.

The group moved to point six of the Quick Reference document that dealt with the number of constructed response items per content area. Mrs. Doorey suggested that a smaller number of constructed response items still has an impact on instruction. Mrs. Woodruff stated that two constructed responses are not enough for validity and reliability. If we want to push to release items, we have to have more than two. Nancy Doorey mentioned that samples could be released. Mrs. Woodruff stated that the RFP must say that we want extended responses combined with multiple choice to get a valid and reliable score and we want to be able to release items as well. If the vendor says that two is appropriate then that is fine. Mr. Sokola stated that if we reduce the number then we save money and time.

Mrs. Corbin made a statement that the task force seems to get into detail then it deviates. We cannot say we are going to be bold then couple that with a conversation about funding. The end result becomes diluted. Mrs. Woodruff responded that “cost effective” does not mean cheap.

Mrs. Doorey stated that if we release a couple of items each year with the new assessment system, then they can be brought into the professional development environment and in the end, instruction can be improved. Mr. Sechler added that he gets the most value from images that he brings up of student writing. This affects the classroom. Mrs. Allen mentioned that if you only offer two constructed responses, how much weight should be given to each. It was agreed that the RFP should specify that when constructed response items are released they should be released along with student responses.

The group moved to point seven of the Quick Reference document that dealt with alignment. The group felt that the growth assessment needed to be aligned to the standards, not aligned to the summative assessment.

The group moved to point eight of the Quick Reference document that dealt with the release of a full version of the test. Mrs. Woodruff mentioned that this would be very expensive. DOE had considered doing that in the past but the cost is excessive. Doing this doubles the cost because new questions have to be developed to replace the old. Mrs. Doorey said that parents feel as if the DSTP is a black box and they want a sample of the full version. Mrs. Wagner replied that we cannot expect the General Assembly to be unrealistic about what can realistically be done. Mrs. Woodruff stated that we cannot release a full version but can release some items.

Mrs. Corbin mentioned that we need a public relations strategy. Some parents need extra support in understanding this process. Mrs. Woodruff stated that DOE does communicate and sends information to schools but she cannot guarantee that it is getting passed to the parents. DOE will continue to communicate with the public. Mrs. Johnson added that some parents want to know what the score means on the DSTP but cannot get down to DOE in Dover to view the test. Districts could scan the test and hold a copy for parents to view. Mrs. Woodruff said that we would need more staff and more security to do that. Testing coordinators cannot do that.

The group moved to point nine of the Quick Reference document that dealt with system oversight and enhancements. Mrs. Wagner stated that we do this now. Mrs. Woodruff said that DOE gets feedback through surveys, meeting with test coordinators, and meeting with vendors. Mr. Sokola wondered if the Option A proposal did not state this because since we already do oversight and enhancements, it is implied. Mrs. Wagner

said there has to be a limit to the amount of change that can happen. Change happens with time. Mrs. Doorey sees this new piece as something that would look at significant changes along the way. Mrs. Woodruff added that we cannot do an RFP every year. Mr. Harter said that if changes are implemented in increments might it be important to designate a stable group to follow and guide the changes. Mrs. Woodruff does not believe an ongoing group is needed.

Mr. Sechler was not sure what “independent” referred to as mentioned in point nine. He mentioned that the task force is independent and the administration gets a lot of feedback from people along the way. Mr. Sokola stated that it is important that he have the public’s input. Mrs. Woodruff offered a solution. DOE gets a lot of input. It could compile a report and bring various representatives around the table to discuss. Mr. Harter believed that this would protect the department. Dr. Roberts added that DOE does a good job with communication but we need to make it broader.

The group moved to point ten of the Quick Reference document that dealt with Vision 2015. Mrs. Woodruff stated that it will take about six months to do the RFP and 2015 does not have a place in this report. Mrs. DiPinto added that the 2015 report has not been released and we are being asked to tie this task force’s recommendation to an unknown. Mrs. Allen added that the timeline would not work. This task force wants to get a recommendation to the General Assembly before they go out of session. She is not sure that 2015 will get to that level of detail before then. Mrs. Woodruff stated that this task force will inform the ongoing discussion in 2015. Mr. Harter also added that things would still need to phase in as well.

Mrs. Doorey asked what would happen if we lock into an RFP then we cannot implement 2015. She wanted to add into the RFP that an information data system be broad enough to deal with 2015. Mrs. Woodruff stated that we cannot do this because it is not a function of the task force to determine the data system.

Mrs. Wagner asked how we get buy-in from the Joint Finance Committee. Mrs. Woodruff replied that we will tell them we will need to phase this in. We cannot roll this out without the infrastructure.

Mrs. Doorey mentioned that Appendix B, C, D, and E were not handouts and should possibly be removed from the final report. Mrs. Doorey also handed out a letter from a teacher of the Springer Middle School.

Mrs. Woodruff closed the meeting. She stated that we will make additional changes to wording in the Option A proposal. The report is due May 31st. Once we get the report to the task force members we will need feedback ASAP. There will be no more meetings.

HJR 4 with SA 2 – Assessment Task Force

Meeting Notes

December 15, 2005 (REVISED)

Location: Cabinet Room, Delaware Department of Education

Time: 8:45 AM

Attendees: Jean Allen, Edie Corbin, Cindy DiPinto, Nancy Doorey, Susan Haberstroh, Bruce Harter, Yvonne Johnson, Martha Manning, Nicole Quinn, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Janine Sorbello, Robin Taylor, Nancy Wagner, Howard Weinberg, Valerie Woodruff

Public: Rhonda Lattin and Elizabeth Reddin

Supporting Documents: Delaware and Federal School Accountability Requirements (Appendix B)

Secretary Woodruff is the Chairperson of the Task Force and provided the agenda. The meeting was called to order and the attendees were given time to review the prior meeting's minutes. Secretary Woodruff mentioned that the meeting is a public meeting and introduced the members of the public that were in attendance. In attendance from the public were Rhonda Lattin and Elizabeth Reddin.

Secretary Woodruff reminded the Task Force of its overall goal. The Task Force is to discuss the current system of assessment, not accountability, and what assessment might look like in the future. The Task Force should consider what recommendations they could offer given the parameters of the No Child Left Behind (NCLB) mandates. Secretary Woodruff reminded the Task Force that the end goal is to help kids, not adults.

An overview of federal and state mandates was given to all attendees. Robin Taylor reviewed each federal NCLB mandate and reported on how Delaware has adopted the mandates within the Delaware School Accountability system. While reviewing, questions arose from the attendees regarding accountability issues such as the calculation of school ratings using Adequate Yearly Progress (AYP) and state progress indicators and transition academies at the high schools versus keeping children in middle school if they do not score at or above Performance Level 2 on the 8th grade DSTP.

Special mention was given to student exemptions from the state assessment. Students are not permitted to be exempt. Students with significant cognitive disabilities must participate in an alternate assessment (the Delaware Alternate Portfolio Assessment) and there is a cap of 1% that can be proficient. There is a waiver provision that states can attempt to seek from the USED if they exceed the 1% cap. Another waiver is given to students who are English language learners (ELL) and have not been in the country for one full year. These students must still participate in the Mathematics assessment but are not required to participate in the English Language Arts (ELA) assessment. There are, of course, a variety of available accommodations for these students in the Mathematics assessment.

There are over 70 foreign languages spoken in Delaware therefore a variety of translators are provided to students who need them to complete the DSTP with that accommodation. The DSTP math assessment is available to ELL students in both English and Spanish, depending on the accommodations provided to those students.

Another aspect of federal regulation is a peer review. Every state must undertake a peer review of their state assessments and standards by the end of the 2005/2006 school year and be approved by June 2006. Delaware will go through the peer review in February 2006. If a state makes a change in their assessment program, the state must

go through peer review again. Peer reviewers are national experts chosen by the United States Department of Education (USED). They consist of both state personnel and experts in particular fields. The peer review teams consist of three people and the USED determines the team without any knowledge from the state as to who is on their particular team.

A Task Force member raised concerns with the true ability of the Task Force to make serious changes. Secretary Woodruff assured the Task Force that their role is important but also reminded them that their role is to make recommendations and those recommendations must be within realistic parameters of change. Task Force recommendations will go to the legislature and the Governor for consideration. Therefore, recommendations should be made with care. Also, any changes to the test will have a direct affect on student test performance because changes in assessment systems usually result in a decline of scores.

A Task Force member requested a timeline for the DSTP, including each decision point related to the current contract to know how program transitions may/can occur over time. The contract with Harcourt, the testing vendor, expires in 2008 and a new Request for Proposals (RFP) will be needed for 2009. A list of TAC (Technical Advisory Committee) members including a brief biography of each member was also requested.

The next part of the meeting discussed the value of Delaware-developed assessment items. Historically, vendors were not adequately prepared to assist the state departments of education in item development. Therefore, a decision with the original stakeholder group charged with the development of the initial assessment system was to create test items through Delaware teacher participation. These test items were to be criterion referenced. Going forward, a question posed was whether the Task Force wanted to recommend that a vendor be hired to create test items or should items continue to be developed through the Department of Education's teacher item developers?

Various pros and cons were discussed. Comments from Task Force members were as follows:

The process of item creation takes time and items must be field tested. The benefit is that Delaware teachers have a vested interest in the items and the items do not show bias to Delaware students.

It does not matter where the questions come from as long as the questions reflect Delaware content standards. However, the teachers that have participated in this process have benefited from the process professionally. This is a professional development opportunity that we may not want to take away from the teachers.

If the Department of Education out-sources the writing prompt the items might not be cohesive with Delaware culture. When the Department of Education's teacher staff writes the items themselves, the items come from Delaware experience and reference.

We should use nationally developed items reviewed by Delaware teachers to ensure the items are relevant to Delaware students.

DOE should not be in the test writing business. There are other things to put time toward.

If teachers like it and are good at it, should the Department of Education not leverage that? We should also consider the impact of taking the task away from teachers and therefore should consider discussing this topic with those in the field. On the other end, these teachers have been trained in assessment development and can return to the classroom to guide other teachers in classroom assessments as well.

It is important to the business community to incorporate communication in writing (not just multiple choice items) to extend communication skills in all content areas so they can be good communicators.

What would the costs be for item development from teachers versus item development from an outsourced company? Currently the contract sets an overall dollar amount but we will try to divide the costs for this. The scoring of the open-ended portions is a significant cost of the assessment.

When only some teachers develop test items then only some teachers are engaged in the process. It would be more beneficial to have all the teachers engaged with assessment results rather than only some in the development phase.

While discussing the layout of the test it was determined that state testing standards measure different things than the SAT does. SAT is a predictor of potential college success.

Teachers would like an assessment that provides teachers with immediate feedback for instructional purposes. Currently the scoring of the open-ended portions for all assessments takes approximately three to four weeks and is done by Harcourt. An assessment with more instructional comments and more immediate results could help in instructional improvement.

Although teachers want a quicker turn around on results, Delaware gets test results fairly quickly compared to other states. Task Force members were reminded that formative assessments may provide more immediate feedback and can inform instruction; however, these types of tests may not be used under NCLB because they do not necessarily test at the grade level. For example, Idaho has a two part process, one for NCLB and the other for teacher feedback. A comment was made that it was probable that the USED would not approve Idaho's hybrid system; however, there were conflicting comments on this.

A committee member stated that both the Business Roundtable and Rodel Foundation have called for an assessment system that measures individual student growth over time on a continuous scale, because they see this as the fundamental unit of measurement that is needed if our state and schools are to have the information required to quickly identify best practices and continuously improve.

Comments were made regarding having an assessment system in Delaware that provides both formative and summative assessments. Optimally, Delaware needs to develop a system, not a single test. The teachers want formative tests and the state and federal mandates require summative tests. Currently we have an assessment that serves dual purposes. In order to be innovative Delaware must think outside of the mandates and look at other states' assessment tools as well. (Handouts summarizing the type of assessment of each state were distributed.) Computer based assessments were mentioned as being the future of assessments.

There were comments regarding the past and the establishment of "national standards." Although this would help in comparing students across the country, the actual development would be difficult because of the idea of local control among states and districts. The National Assessment of Educational Progress (NAEP) is currently the only national comparison assessment and is limited in its ability to show true comparisons because it is only a "sampling" of students at certain grades. It was mentioned that Delaware is leading the nation on improvement in reading.

A question was raised about student accountability mandates. More information was requested regarding this topic.

A request was made that representatives from North Carolina and Virginia speak to the Task Force about their high school end-of-course exams.

A request was made that Brian Gong and Bob Linn be brought in to talk to the group and also possibly someone with knowledge of computer-based and computer-adaptive assessments.

During open discussion the business community again reminded the Task Force that they need to focus on overall communication skills, not just math and science. There was also brief discussion of possible end-of-course assessments.

The public commented that cost is the most important aspect in regard to test creation because money is needed to go toward other social programs within the state.

Overall, Delaware needs to look at growth over time, resource allocation, a multi-assessment system, cost, and a scale by which Delaware can be measured nationally. There has been an effort to create a common scale with state assessments but this is still far off.

Upcoming meetings were scheduled as follows:

January 12, 2006 at 8:30AM

January 31, 2006 at 1:30PM

February 10, 2006 at 8:30AM

Meeting adjourned at 11:00AM.

HJR 4 with SA 2 – Assessment Task Force
Meeting Notes - REVISED
January 31, 2006

Location: Cabinet Room, Delaware Department of Education

Time: 1:30 PM

Attendees: Jean Allen, Vicki Cairns, Cindy DiPinto, Nancy Doorey, Susan Haberstroh, Bruce Harter, Yvonne Johnson, Martha Manning, Nicole Quinn, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Robin Taylor, Nancy Wagner, Valerie Woodruff

Presenters and Members of Public: Ellen Forte, Brian Gong, Elizabeth Reddin, Rhonda Shulman Lattin

Supporting Documents: Assessment Task Force Statement Draft (Appendix C)

Secretary Woodruff, Chairperson of the Task Force, called the meeting to order and introductions were given from all attendees. In attendance from the public were Rhonda Shulman Lattin and Elizabeth Reddin. Secretary Woodruff distributed the prior meetings minutes, a handout on state exit exams, and a handout on state promotion and retention policies. She reminded the task force members that these documents, plus additional documents, were sent to each member via email, with the exception of the exit exam information.

Secretary Woodruff asked that the two guests in attendance, Ellen Forte and Brian Gong, introduce themselves and discuss their background, experience, and views on assessment. The Secretary encouraged the task force to ask questions.

Ellen Forte introduced herself as an independent consultant who has worked in a variety of positions and at a variety of agencies. These positions include Assessment Director for Baltimore City and peer reviewer for the U.S. Department of Education (USED). The agencies in which she worked include Connecticut's Department of Education, edCounts, the United States Department of Education, among others.

Brian Gong joins us from the Center for Assessment, located in New Hampshire, where he is the Executive Director. Among his experience is that of USED peer reviewer and he is an expert in comprehensive assessment systems.

Dr. Forte, being the first to present, commends Delaware for being forward thinking in the area of revising assessment systems, noting that this forward thinking approach is rare. She emphasized that every state must start with content standards. The assessments must be driven by standards. She stated that the assessment must be aligned with our standards. She stated that the assessment doesn't drive the system, but that the assessment is driven by the standards.

To assist states with compiling evidence for the standards and assessments peer review process under No Child Left Behind (NCLB), the USED has created a document called Peer Review Guidance. This publication addresses a number of topics including but not limited to standards, alignment, inclusion, alternate assessments, and score reports. Dr. Forte indicated that the technical requirements covered in the Peer Review Guidance are driven by the "Standards for Educational and Psychological Testing" (jointly published by American Educational Research Association (AERA)/ American Psychological Association (APA)/National Council on Measurement in Education (NCME).

Dr. Forte reiterated the requirements of NCLB and emphasized that in NCLB all students must be included in the assessment process, including students with severe cognitive disabilities. The alternate assessments for these students must also be aligned with the state's content standards.

Score reports must also be aligned with expectations. The state is responsible for producing a meaningful score report for each student that helps the parents understand what the expectations and the goals are within the student's education.

Dr. Forte then discussed the peer review process. All states must undergo a peer review process for their standards and assessments between February 2005 and June 2006. The state must submit evidence of their compliance with the standards and assessments requirements of NCLB to be judged by experts. To date, 28 states have gone through the review process; no state to date has received full approval or full approval with recommendations. These states have been asked to provide additional information or evidence. She said that because many states are in a transitional period with their assessments and not all states have had assessments in place at all the required grades, it is not surprising that states are not getting full approvals. If a state makes substantive changes to the standards or assessments the state must undergo the peer review process again.

As Delaware goes through the peer review process, the department will receive communication throughout the entire process. As our information goes through the various channels of peer review, feedback will be given. The peer review team will make a recommendation on whether approval should be granted fully or with conditions, but the final decision lies with the USED.

Dr. Forte was asked which state has similar demographics to Delaware and if there was a specific state's model she would recommend for Delaware. She mentioned Wyoming as an example for their innovative approach to assessment. Wyoming has a new assessment this year that is designed to meet NCLB requirements and at the same time give teachers the feedback that they desire by offering a mixed use test. Their reporting method has also been innovative, using a teacher to teacher approach that offers recommendations. Their system involves both technology and immediate feedback (on a portion of the assessment).

Dr. Forte was asked if one assessment can do everything. She responded that it is nearly impossible for a large-scale statewide test to do everything; the results can be used for specific purposes and then combined with other information to make additional decisions. Dr. Gong responded that theoretically it is possible but he has not yet seen a test that does many things well. Oregon's statewide assessment and the Northwest Evaluation Association's (NWEA) Measures of Academic Progress are both adaptive tests. Even though fully adaptive assessments can yield more accurate individual scores within a shorter testing time, they do not meet NCLB's requirement of on-grade testing. Oregon's is adaptive within the grade level only, so it can become somewhat easier or harder based on the students' answers, without moving to the next grades' content.

At this point, Brian Gong began his discussion stating that assessments take a long time, normally two to three years, to bring things together operationally, including planning, development, field testing, administration, and standard setting. He addressed three points:

1. Validity is in the interpretation and use of the test, not the test itself. Tests are designed for a specific purpose. For example, a school would not use the SAT for an end of course exam. You first must define what you want the test to do. A final exam might be designed to include cumulative information from the entire class or just information from the last half of the class. That depends on how the teacher defines "knowing" the information. How the test is administered also requires a lot of thought.
2. Theory of Action: "How am I going to use the test?" and "What difference will it make?" NCLB is a minimal test, testing only reading and math. The focus of NCLB is to bring the bottom up, as it gives no credit for raising achievement above the Standard. He talked about the differences between a criterion-based

assessment and a norm-referenced assessment. Dr. Gong discussed a Comprehensive Assessment System that entails different roles and responsibilities. The state has certain roles and responsibilities while the district, the schools, the parents, etc. have others. One state test is difficult to fulfill all these various roles and responsibilities. Most state tests are designed as “end-of-year” and, therefore, will not have as much detail for diagnostic purposes. If an assessment is to be diagnostic and for use by teachers throughout the year then it needs to be given near the time of instruction.

3. An assessment system is always evolving. Dr. Gong stated that testing in education is messy and complex. He indicated that different tests are differentially sensitive to learning.

A question was asked about the measurement of achievement against the standard versus measurement of individual student growth. Dr. Gong responded that he believes individual growth is a better measure for accountability and that he had argued with US DOE that NCLB should focus on individual student growth rather than cohort status, but that he had “lost that battle.” The task force member then asked if Dr. Gong felt the task force should consider, as the next-generation DSTP, a two-phase computer-adaptive system in which the first phase is adaptive within grade level and the second phase fully adaptive to get the more accurate individual score and growth. Dr. Gong said this would be a reasonable model for the group to consider.

A question was asked about the National Assessment of Educational Progress and how the comparison is made between states when there are no state standards. Dr. Forte responded by saying states are required to have rigorous content standards and assessments. Content standards and performance standards are two different things.

One task member asked about a hybrid system, one that measures against the standards and also measures individual growth. Dr. Gong responded by asking if it was important to have one test to measure both of those things. If we know what we want to do, then we can get the experts in to tell us how to do that. Dr. Gong stated that we have to start with decisions, then build a system by choosing tests wisely depending on our purpose(s). We can streamline but we have to figure out which tools are available for the purposes we want. With instructionally sensitive assessments there still may be an issue of turnaround time for the results.

Dr. Gong mentioned Washington state’s investment in professional development and assessment literacy (e.g., on use of the assessments) so that teachers are able to create and use assessments. He also mentioned the Regents exam in New York. Since it is scored locally there is a fast response with the results. However, it was mentioned that this sometimes causes a credibility problem. We have to look at the reasons why we make our choices in order to be innovative. The NWEA exam was discussed and it was noted that there are benefits to this type of exam, but one of the drawbacks is the test length. [Is this right? I have no notes about test length being discussed – as it ends up being the same number of items as DSTP!] The test needs to be long enough to ensure that the standards are being measured. A test may be reliable but not necessarily covering all the content. Computer adaptive tests may solve some problems but may not have content coverage. Oregon was mentioned as an example of a state that has a computer-administered, adaptive test. It was noted by Mrs. Doorey that Oregon assesses several times each year, supplements the computer-adaptive assessment with student-constructed response items that are scored within three weeks, and that scores can be “banked” through the year.

Secretary Woodruff addressed the issue of how we can look at different ways for students to present what they know. Rhode Island has shifted assessments to the classroom and is requiring that classroom evidence, end-of-course exams, common tasks, and exhibitions be incorporated. Comparability between states was brought up as a concern.

A comment was made that we need to keep the students with disabilities and English Language Learners in mind as we move through the conversation. How do we measure what they know and how they have improved? This holds

true for those students performing below grade level. Comments were made that these students may not meet the standard but may be improving and the assessment needs to show that.

Secretary Woodruff mentioned that the Delaware Department of Education is working on a recommended curriculum. English Language Arts and mathematics will be completed by summer 2006. The districts are required to show that their district curriculum is aligned to the standards. Eventually the statewide recommended curriculum may make it easier to use end of course exams and could be a part of our system to help students improve. Mrs. Woodruff reiterated that we are looking at a system and not necessarily one test.

A question was asked whether we can have an assessment for more than one purpose. Dr. Gong said that in general it is very difficult. He commented that states have put 90% of energy into meeting the requirements of NCLB but not on being educationally strong or useful, and that the primary purpose of NCLB is to allow states to detect if there is a problem in the school, not student level information. States need roles and responsibilities, a comprehensive system, to do that. In a comprehensive assessment system the state needs an assessment and needs to invest in professional development so that professionals can interpret the assessment.

A question was asked if the focus of NCLB has changed the goals of assessments. Dr. Gong responded by saying that assessments are getting less complicated across the states, primarily because of budgets. Test development and scoring cost of extended response questions cost the most money. Dr. Gong encouraged Delaware to step back and think about the educational purposes of the assessment.

A comment was made about the NWEA test. This type of test has some good features and provides immediate feedback to teachers. However, one of the districts using this assessment noted that it is not necessarily cheap (inexpensive).

A comment was made that the structure of a test (computer vs. paper) can affect the performance of a student. Dr. Gong stated that there has not been enough study in this topic and it is hard to compare the validity of scores between a computer-administered test and a paper-based exam. This question of whether the test is measuring knowledge or format constraints (scrolling on the computer, moving between paper booklet and response form) is true for every type of test. He also noted that the correlation between two versions on the same paper-and-pencil test is about .90, which means that 19% of the schools are given erroneous accountability ratings.

Secretary Woodruff announced that Ellen Forte will be attending the next meeting and that the focus should be on establishing our purpose.

Upcoming meetings were scheduled as follows:
February 10, 2006 at 8:30AM

Meeting adjourned at 4:20 PM.

HJR 4 with SA 2 – Assessment Task Force
Meeting Notes -REVISED
February 10, 2006

Location: Cabinet Room, Delaware Department of Education

Time: 8:30 a.m.

Attendees: Jean Allen, Vicki Cairns, Edie Corbin, Nancy Doorey, Susan Haberstroh, Bruce Harter, , Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Robin Taylor, Nancy Wagner, Valerie Woodruff

Guest: Ellen Forte

Secretary Woodruff called the meeting to order and circulated the agenda. She thanked Ellen Forte for coming to the meeting and said that it is important to have Dr. Forte with us because she has the national perspective. The first agenda item was a review of the last meeting. Mrs. Woodruff reminded the members that we had discussed the landscape of assessment and had heard the national perspective as to the “have to have” versus the “nice to have” elements of an assessment system. The question raised is how to mesh the two. Mrs. Woodruff commented that we talked about a system rather than one assessment.

A task force member urged that the task force place highest priority on a system that will help improve teaching and learning, and ensures that it also meets requirements of NCLB. Mrs. Woodruff agreed.

Mrs. Woodruff commented that there had been requests for scheduling experts to come and talk to the committee. Nancy Doorey suggested Ron Hambleton. Robin Taylor has arranged for Dr. Hambleton to participate at the next task force meeting by telephone since he is unable to attend in person. The next meeting is scheduled for February 27 at 1:00 p.m. Dr. Hambleton will be available at 1:30 p.m. Mrs. Woodruff stated that we are still working on getting other people from other states to come in to talk about end-of-course exams.

Mrs. Woodruff commented that today’s meeting was to focus on the purpose of the assessment system. She posed the question “What do you think the purpose of an assessment should be?” to the task force and asked the members to take a few moments and write down their thoughts in a sentence or two. The following was reported:

Dorothy Shelton – to fairly and accurately communicate student achievement and inform instruction while fulfilling national funding requirements

Dave Sechler – inform student learning and instruction; measure achievement against the standards; meet legal mandates

Nancy Wagner – to move forward the education of the individual student while satisfying federal requirements

Dave Sokola - to have the necessary components to measure student, teacher and the system - performance and progress

Nancy Doorey – 1) provide detailed information at regular intervals to teachers, parents, and students about each child’s instructional needs and progress relative to the state standards; 2) provide the public with a fair evaluation of academic growth of each school and district relative to the standards; 3) provide educators with information to tie student growth to strategic management decisions

Vicki Cairns – inform instruction; measure student progress toward individual goals; identify achievement against reliable standards, not as an evaluation of teacher performance

Edie Corbin – provide a process within which the state, local educators and other constituencies can measure the impact of comprehensive instructional practices and methods and then to evaluate and assess student academic progress toward a set of standards.

Bruce Harter – provide information for decision making that improves student learning.

Nancy Wagner made the comment that we need to capture the thought from the last meeting that an assessment tool cannot do everything. We need to ask what the DSTP is capable of doing. Mrs. Woodruff stated that we have used the DSTP for various purposes, some of which may not have been appropriate.

Mrs. Doorey stated the task force should consider this to be a golden opportunity for the state to do something really good. We need to be thinking about the ultimate goals of an assessment rather than limit our thinking. She passed out copies of the Education Policy Statement of the Delaware Business Roundtable and the Delaware Chamber of Commerce, as well as the results of a survey of the DSBA Board of Directors (copies attached). Mrs. Woodruff agreed, but suggested that we need to be cognizant of what we can do and can afford. Mrs. Woodruff stated that we need to strike a balance between what can be done and the cost. She also said that because of the assessment environment our next system will be more expensive.

A comment was made that we need to build a compelling vision of what the assessment system can do and then we can build support. The task force recognized the need for thoughtful discussion about the next assessment system.

A comment was made that we need to be cognizant of the two bills that were introduced in late January. A discussion ensued regarding the politics behind the legislation and the public's support for changes to the DSTP. The task force agreed that there needs to be a communications strategy. This communication strategy would require all stakeholder groups represented on the task force, not just the Department of Education, to communicate the message that there is movement in looking toward changes to the assessment system, but the movement needs to be thoughtful and not instantaneous.

It was suggested that perhaps we need to enlist the assessment experts or one of the officials from NWEA to get a message about the current status of the growth pilot. Mrs. Doorey said that she will be presenting an overview of the NWEA pilot to the House and Senate Education committees when the General Assembly reconvenes, and offered to work with the task force and Secretary Woodruff on a message about the harm of a bill that would require an immediate change in the state assessment.

David Sokola made a comment about the previous evening's meeting held in the Brandywine School District with parents of special education students, in which there was widespread desire to transition to an assessment like NWEA's. It is important to keep in mind these students and parents as we move forward since the current DSTP is not always well received by the parents of special education students. A critical mass of parents is important for any assessment to be accepted and maintained. We need to give the public a system that they want. Wendy Roberts mentioned that there are some things in the works at the federal level for the special education population but the final guidance has not been released.

Dr. Forte mentioned that there was an article in Education Week last week indicating it is a dangerous game to move too quickly on these types of changes. We need to be deliberate. It takes 3-4 years to put a system together and then several years after that to get the data to see if the system is working. Mrs. Woodruff concurred by stating that the successes we are seeing with our students today did not happen overnight. Comments were made regarding divided loyalties and what is right versus instant gratification. Ms. Corbin suggested we leverage the voice of the community and grassroots organizations by getting them to understand the time it takes to make changes to our assessment system.

Bruce Harter stated that we need to have a collective voice and have a communications plan from this group to help shape the expectations of the public. Mrs. Allen reminded the task force that the DSTP is not the first Delaware state test.

A comment was made regarding the test effect, stating that a change in an assessment may result in decreased performance simply because it is a new test. A comment was made that we need to discuss what we have and should keep and determine what we don't have and need.

Dr. Forte reiterated the difference between an assessment system and an assessment tool. We need to determine what tier the information from an assessment will be used. The state needs information on a different level than the district, school or classroom. There is a distinct movement to looking at a coherent system of assessment. The ideal would be more connections between state and local assessment systems, all linked with the standards. A statewide assessment may provide information on a programmatic level as opposed to the classroom level. The state assessment is a snapshot and directs the state's discussion on where resources should go. Teachers need more of a mile deep look whereas the state assessment is more of a mile wide perspective.

There was also a discussion related to the need for more teacher assessment literacy.

Mr. Sokola mentioned that when there were no (student) consequences to the assessment; that kids just "blew off the test." A conversation ensued about how we need to think about and build a system that deals with this issue of motivation.

There was further discussion on how we need to be clear on what the current DSTP can and cannot do. Mrs. Wagner suggested that we have taken a statewide test and we have tried to use it for other things. She suggested that this is what the general public has voiced their displeasure about.

Dr. Forte discussed the assessment being instructionally sensitive versus instructionally supportive. She indicated that a large-scale assessment can inform decisions on a grander scale (at the school and curriculum level) but cannot be instructionally supportive. Attention to alignment of the assessment to the standards can get at instructional sensitivity of the assessment. She noted that it is important that once we change the purpose of an assessment that we are able to provide evidence that the assessment is valid for the purpose.

Mrs. Woodruff commented that the assessment is used to measure state, district and school accountability as well as student accountability. The student accountability uses of the assessment were added as an additional purpose. The initial intent was for system accountability. During the original discussions there was a desire to do embedded assessments that would have provided more in depth knowledge of student information, but it was too expensive at that time.

Mrs. Doorey commented that items in a comprehensive system could include a statewide assessment that would be summative and considered high stakes; benchmarks that could provide measuring individual student growth toward the standards over time; and optional assessments measuring attainment of the learning goals of short instructional units. There was discussion regarding an item bank. Comments were made that teachers need to keep in mind what they are doing throughout the year regarding assessment and that we should think about an investment in assessment literacy. Mrs. Woodruff commented that there would be model assessment items in the Statewide Recommended Curriculum.

Comments were made on the need to inform the public and other stakeholders that it is good practice to review assessment systems, as the task force is doing, and that the DSTP is not a "failure." Mrs. Woodruff stated that her office will put together a statement outlining where we are as a task force. She asked the members to assist in pulling together the important concepts that were outlined during the earlier part of the meeting. She reinforced the comment that we need to come up with a communication plan that informs the public that the task force exists and that we are being thoughtful and deliberate as we move forward. Vicki Cairns commented that the teachers are aware of the task force through their DSEA Alerts.

Mrs. Wagner asked Dr. Forte if there are any states that use the system that we are discussing. She stated that no state has all the elements we are looking for and that states are now only beginning to have the conversations about assessments, similar to what this task force is doing. Dr. Forte mentioned that the West and Northwest have been using technology mostly because of the rural nature of the states. She mentioned that there has been great growth in technology and assessment in general. Comments were made that Delaware may find the need to pull different pieces from different states to create the assessment system that is desired.

The meeting was adjourned and Mrs. Woodruff stated that the “statement” would be developed and shared with task force members.

Next meeting: Monday, February 27, 2006
1:00 p.m. – 3:30 p.m.
Townsend Building, Cabinet Room

1:30 p.m. - Telephone Conference call: Ron Hambleton, Ph. D., Professor, Chairperson of the Research and Evaluation Methods Program, Co-Director of the Center for Educational Assessment, University of Massachusetts at Amherst

3:00 p.m. - Telephone Conference call: Lou Fabrizio, Ph. D., Director, Division of Accountability Services NC Department of Public Instruction

HJR 4 with SA 2 – Assessment Task Force
Meeting Notes
February 27, 2006 REVISED

Location: Cabinet Room, Delaware Department of Education

Time: 8:30 a.m.

Attendees: Jean Allen, Vicki Cairns, Cindy DiPinto, Nancy Doorey, Emily Falcon, Susan Haberstroh, Yvonne Johnson, Martha Manning, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Robin Taylor, Nancy Wagner, Valerie Woodruff

Members of public: Fred Bost

Secretary Woodruff called the meeting to order and mentioned that the Department had submitted the state's proposed growth model to the USDOE for state accountability purposes. Twenty states submitted plans (14 for consideration this year and 6 for consideration beginning next year) while the USDOE was proposing allowing up to ten states to participate in this pilot.

The agenda, additional copies of the meeting notes for the January 31 and February 10 meetings, and a draft "statement" were made available to those task force members who needed copies (the items had been attached to a previous email). Mrs. Doorey commented she had changes to the meeting notes and Mrs. Woodruff stated that the meeting notes would be revised appropriately.

The draft "statement" was discussed and rather than prolonging the conversation, Mrs. Woodruff stated she would revise the statement based on the conversation and send out a revision to the members for review. Many of the comments revolved around the clarification that we are talking about a "system" and not necessarily one instrument. The purpose of this statement is to provide a foundation or reference point for the task force and then have the pieces of the assessment system cascade from there.

The second item on the agenda was to have a conference call with Ron Hambleton, Ph. D., Professor, Chairperson of the Research and Evaluation Methods Program, Co-Director of the Center for Educational Assessment, University of Massachusetts at Amherst. The task force had expressed a desire to have someone with expertise about computer based assessments. Dr. Hambleton is a member of the DSTP Technical Advisory Committee.

A question was asked about what least desirable effects, if any, there are for students taking computer based assessments. Dr. Hambleton opened by qualifying his remarks to say that all of his answers were "off the top of his head" and he reserved the right to come back in follow-up. He said that there is the issue of how well the assessment works. Paper and pencil tests do not always test the same way as a computer-based test. Reasons include scrolling, lack of familiarity with the computer; some things test better, some not as well. He gave an example of students writing using the computer, and the requirement that in some formats, math for instance, that there is more of a need to copy the computer question to paper. He also mentioned if there is a choice of taking the test on computer or paper/pencil then there is a good size equating problem; cut scores set on a paper/pencil test may not always translate to the same cut score as a test given via computer. He said if you use both there is an issue of making the scores comparable. He did mention that Virginia and North Carolina are moving toward computer tests.

Dr. Hambleton indicated that scores come back to the teachers more quickly, but typically this requires a move toward more multiple choice questions. To do any other type of questions with automated scoring would take

elaborate software. He did not think that moving toward more multiple choice questions was necessarily a good thing because you can't span the broader curriculum. He does support moving to computer based but there must be the capability to score the assessment and still measure higher level thinking skills. He said you need to have constructed response questions. He said that there are some professions such as architects and accountants that use computer based assessments for credentialing; however, these require extensive testing and are expensive. He noted that to get comparability you need to do some extensive equating studies. Dr. Hambleton said there may be some risks for some subsets of students in the short term by using computer based assessments.

Mrs. Woodruff explained to Dr. Hambleton that we are in a "fact finding" mode. He stated that he thinks some of the educational testing vendors will be where some of the credentialing organizations are in about 20 years.

Mr. Sokola mentioned the recent articles in Education Weekly highlighting Delaware's performance on the NAEP and that constructed response questions were given a high value. He said that he envisioned a hybrid with adaptive components and then some constructed response items to get a faster response (scoring) time for a portion of the assessment.

Dr. Hambleton said there are at least 4 test designs using computers:

1. fixed set of questions, maybe multiple versions that are equated and then randomly given to students
2. bank of items and each students sees different items
3. "blocks" or "testlets" that move to a harder or easier block depending on student's responses
4. adaptive which changes to harder or easier after each question

He mentioned that there may be a potential issue of security if all students don't take the test at one time and that the constructed response questions would have to change if not given at a common time. Creating different constructed response questions and developing rubrics would be expensive and there are issues of equating. Mrs. Doorey suggested cycling the kids through the computer at the same time and then doing the constructed response at a later time with paper and pencil tests. Dr. Hambleton suggested this may work if you have a large enough bank of items or give the assessment to different grades on different days. His biggest concern was protecting the security of the constructed response items. There was conversation about how the students may react. He said that Massachusetts provides early reports to the multiple choice portion of the assessment (given on paper/pencil) while the reports on the rest of the test come later.

Mrs. Doorey referred to Dr. Hambleton's co-edited book and said there was an argument for using more computer based assessments because this is what the future will be.

Dr. Hambleton mentioned that he was concerned with teachers' knowledge of assessment and would suggest working on providing more pre-service and in-service professional development on this area. He also would work on providing better score reports. He suggested a focus on a 2 tiered system and doing what is needed to satisfy federal requirements and state accountability and then have a second, more flexible part that could provide more diagnostic information to teachers. This second part could be computer based or computer adaptive.

He complimented the work by the state on the alignment of our standards and grade level expectations, development of performance level descriptors, and revisiting the performance level cut scores. He said many times the negative attitudes heard about state testing are because of a lack of understanding of the purpose. He said he agreed with the comments made by Brian Gong that one test can't do all things. He said that the College Board has given up on trying to provide detailed student diagnostic information on the SAT test. He suggested that there is a limit to the amount of diagnostic data you can get from a state summative assessment.

Mrs. Doorey commented that if the NCLB law lifted the on-grade assessment requirement, couldn't you get more diagnostic information. For instance, use a computer adaptive testlet and then take a multiple choice section as the state test.

Dr. Hambleton likes the concept of computer adaptive testing and said that he is inclined to keep computer based testing separate from the state (summative) test. He thinks Delaware and Massachusetts are doing a pretty good job of measuring student growth.

The next telephone conference call was held with Lou Fabrizio, Ph. D., Director, Division of Accountability Services, Laura Kramer, Chief Psychometrician, and Mildred Bazemore, Section Chief of Test Development at the North Carolina Department of Public Instruction to discuss end of course assessments and the state's testing program. Robin Taylor had called Dr. Fabrizio prior to the conference call so that they could be prepared to talk about a few questions. The questions included - How are the end-of-course assessments developed? How are the results used? How are the results reported?

Dr. Bazemore explained that their assessment program began in 1986, with planning beginning in 1984. Currently there are 10 end of course exams; English I, Algebra I, Algebra 2, Geometry, Biology, Chemistry, Physics, Physical Science, Civics and Economics, and U.S. History. All are multiple-choice items and administered within the last week if block scheduling and within last 2 weeks if it is a year long course. How they are counted has evolved over time from requiring the courses to be part of the final course grade to requiring that the tests count for 25% of the final grade. During the last several years, North Carolina's State Board of Education (SBE) has changed the exit standards and students must be at or above proficiency on 5 different end of course exams, including Algebra I, English I, Biology, U.S. History and Civics and Economics. For the students entering high school in the 2006-07 school year, there will be a graduation project requirement. This is still being developed.

Some of the challenges that were highlighted included reporting student results in the necessary timeframe when grades are being determined; the extent to which the scores count and student motivation; moving to block scheduling; field testing and changing of curriculum. Because the tests are all multiple choice they have been able to get the scores back quickly. English 2 was eliminated because it was an entirely constructed response assessment.

The next question revolved around how the assessments were developed. Most were developed with internal staff and with folks from UNC-Raleigh. They are also working with some testing companies to develop some of the test questions. The State does the administration of the assessments, i.e., publishing and sending the assessments to schools. The assessments are scored locally in the central offices of districts. They all have laser printers, software, scanners, etc. They developed a means by which they pre-equate the tests, so schools can immediately generate and print the Individual Student Reports, class, school and district reports, with the data sent to NC DPI where it is subsequently certified. Dr. Fabrizio responded to a question regarding accountability and said that the end-of-course assessments are used for accountability in the high schools (details on this appear on their web site).

Dr. Fabrizio responded to a question as to what parents thought of the system. He said that the SBE did focus groups and there are numerous safeguards in effect for students. No decisions related to graduation are made solely on a one-time test. Some of the safeguards include the ability to take the tests multiple times; there is a standard error of measurement built into the scoring; if the student still did not meet the standard then there is an external review committee looking at other information; and the final decision of whether a student has met the standard is made by the school principal.

Dr. Fabrizio responded to a question of whether there are incentives for high performing schools. He said there is a hierarchy of classification of high schools: honor school of excellence, school of excellence, school of distinction, school of progress and school of low performance. If growth is met some staff receive monetary bonuses.

Dr. Fabrizio said there are 3 parts to their accountability system: AYP, Growth and Absolute. North Carolina asked U.S. DOE to use English I and Algebra I as part of AYP. The USDOE did not agree initially because some take Algebra I before 10th grade and not all kids take these tests. More recently USDOE said they could use these if there were alternative tests for the students that did not take these assessments. Also, US DOE is requiring them to use the 10th grade writing assessment as part of their ELA score.

Dr. Fabrizio said that there are 4 different courses of study that lead to the same diploma: College – University Preparation; College-Technical Preparation, Career, and Occupational (includes students with IEPs).

Dr. Fabrizio responded to questions related to cost. They have a relationship with the higher education institutions in their area and much of the development is done by these institutions under the direction of NC DPI staff. The cost for the tests is about \$200,000 per year per end-of-course test, plus staff time. They mentioned there are approximately 20 people involved in test development (10 in development and 10 on the operational side) who spend approximately 25% of their time on the end-of-course exams. They also have to pay for usage of any test development done by contractors. The scanning, scoring and printing equipment that they place in districts costs roughly \$12,000 per 25,000 students, and they replace one third of it every year. A request was made to have them provide a spreadsheet of their costs. Costs included staff time, copyrights, printing, shipping, and technology infrastructure. They did not know how much it would cost to have a vendor do the work they are doing along with the local higher education institutions. They also mentioned a high turnover of their staff because of the local companies pulling the DPI staff for higher salaries in the private sector.

The staff at NC DPI also work on end-of-grade exams in the lower grades including reading and math grades 3-8 and writing in grades 4, 7, and 10. The reading and math assessments are typically given during the last week of the school year. Scoring and printing of reports are handled the same way the high school end-of-course exams are handled. They also have an online computer skills test at 8th grade, in the past they have had high school competency tests, 3rd grade pre-tests and alternative assessments. The staff has generated 75 alternative tests (checklists, portfolios, modified tests, etc.).

A question was asked whether the state has a balanced budget. He said yes, but did not know how long that had been in effect. Other responses to questions included noting that school districts set their own calendars for assessments within a window provided by the state, and that the state has had to put in some checks and balances so that students were not moved out of courses if they were not performing well.

Other issues that NC has experienced include complaints by school counselors that there is too much testing and that they are not able to do the counseling they think is needed. Also the amount of testing is compounded by district benchmarking testing; the public sees testing as testing and doesn't make a differentiation between state and district testing. There was discussion regarding the equating of forms.

Dr. Fabrizio responded to additional questions about the graduation project requirement. He said that his division is not playing a major role in this, but rather the curriculum area of NC DPI. In concept he said the project includes picking a topic, writing a research paper and making a presentation to a team of reviewers.

There were additional questions regarding the percentage of staff time that was devoted to end-of-course tests. He responded that for about 20 people it is approximately 25 to 30% of their time on this. The tests are criterion referenced. Parents know how their students are doing in relation to other students in the state, but not nationally. There are approximately 1.4 million students in North Carolina. He said the department spends approximately \$9 million per year on all assessments. It was unclear if this included staff time, so Robin Taylor agreed to request a

detailed breakdown of North Carolina's assessment system costs and, to the extent possible, to put Delaware's costs side-by-side.

Other responses included that the federal approval under NCLB will hinge on whether the state can show that the NC tests are aligned and that the multiple choice questions test their standards. A committee member asked if the US DOE has raised any concerns about their use of solely multiple choice items given that their standards call for high-order thinking. Mr. Fabrizio replied they already had their assessment system peer review and this was never raised as a problem in the Decision Letter from the US DOE; however, the NCDOE does not yet have final approval.

The next two meetings are scheduled for Tuesday, March 7th at 8:30 am in the Collette Building and Wednesday, March 15th at 8:30 am in the Cabinet Room at the Townsend building. Items on the next meeting agenda included a request for costs from other states that do end-of-course exams. Mrs. Woodruff suggested that the next meeting also include a brief summary of the technical aspects of assessment.

HJR 4 with SA 2 – Assessment Task Force

Meeting Notes

March 7, 2006

Location: John Collette Education Resource Center, Delaware Department of Education

Time: 8:30 a.m.

Attendees: Jean Allen, Cindy DiPinto, Nancy Doorey, Susan Haberstroh, Bruce Harter, Yvonne Johnson, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Robin Taylor, Valerie Woodruff

Members of Public: Fred Bost, Charlene Tucker

Supporting Documents: Delaware Student Testing Program (DSTP) Costs for the 2005-2006 School Year (Appendix D) and Standards and Assessment Peer Review Guidance (Appendix E)

Secretary Woodruff called the meeting to order and handed out the agenda (attached). The draft statement was delayed for discussion because the revisions had not been completed. Mrs. Woodruff indicated she would send out the revision before the next meeting. Several other documents were handed out including excerpts from the *Standards and Assessment Peer Review Guidance* (the introduction pieces from each of the 7 sections), and two sheets related to DSTP costs (attached).

Robin Taylor discussed the technical aspects of testing. The technical aspects, as outlined in the *Standards and Assessment Peer Review Guidance*, are the same for any assessment. Ms. Taylor framed the discussion in light of why technical quality is important. In the past, assessments used to come with technical reports and studies since many states and districts were using off-the-shelf assessments; this was typically part of the vendor's responsibility. In the early to mid '90s, there was a move toward hybrid tests or customized development of assessments (such as the DSTP with Delaware developed items). States using customized assessments were responsible for getting the evidence of technical quality. Thus, any change in the construct of a test requires a new look at the validity and reliability of the test.

Ms. Taylor explained the concepts of reliability and validity. Reliability refers to the stability of the test or the extent to which the test scores are consistent. For instance, the same test to the same student equals the same score. The longer a test is, the more reliable it will be. Reliability is a necessary part of validity, that is, with any version of the test providing the same score. Equating is done when there are different versions of a test.

A question was raised about whether there is a correlation between SAT9/10 percentile rank and DSTP performance level. There is not a direct relationship. There is a norm referenced part of the test (first 30 multiple-choice items for reading and mathematics which comprise the SAT10) that provides a percentile rank which gives an indication of how a student compares to the performance of students nationwide; however, the DSTP performance level is based on the entire test which includes the norm-referenced portion along with additional Delaware-developed items. In addition, the entire DSTP including the SAT10 norm referenced portion of the test is based on Delaware content standards. A small number of SAT10 items are excluded from the total DSTP score/performance level because they were determined not to be directly aligned with Delaware's content standards.

A question was asked about the use of confidence levels in the Adequate Yearly Progress (AYP) determination. Confidence levels are used for school accountability because there is always a slight measure of error for each test. When several tests are aggregated for school accountability the effect is larger.

The reliability coefficients for the DSTP typically range in the .80's and .90's. There are several types of reliability. For the DSTP, we look at the internal consistency of our assessments to get a measure of the reliability; for each year a technical report is created which contains the reliability coefficients as well as other technical documentation. Ms. Taylor commented that she was trying to stay as general as possible and not get steeped in "psychometricianese".

Ms. Taylor then gave an overview of validity. She explained that this is the extent to which the test does what is intended, that is, the purpose. There are several different types of validity. Content validity is more than the alignment of the test to standards and looks at the extent to which the content is a match between what it taught and tested. DOE looks at the individual items and checks for the match to the standards. The vendor did an initial match for the SAT10 items to the content standards but we do our own independent check of this alignment with panels of educators; this provides evidence of content validity. We field test all of our items before they are used on an operational test.

Criterion-referenced (empirical) validity is the extent to how well the test does what we want it to do. Predictive validity is how well it can predict. Rater (scorer) validity is the evidence that there is inter-rater agreement. We also look at what is called the accuracy of decision consistency in regards to the DSTP performance levels. DOE compiles a great deal of evidence which then goes into determining validity.

Ms. Taylor talked about rater (scorer) reliability and using anchor papers embedded in the tests the raters score. We look for rater (scorer) drift. Delaware requires all scorers to have a college degree and go through an interview and vendor training. They need to qualify and then remain qualified. Ms. Taylor said that Delaware is very rigorous in the quality control and scoring of our assessments.

Ms. Taylor addressed the importance of the technical reports. She explained that since the mid '80s states are using assessments for different purposes. In the past, there were the basic skills tests used to assess students very generally. Today, this is different. Tests are used for high stakes, they provide information to the parents, are used for student, school, district and state accountability. Because of this, more technical studies need to be done. Also, since 1999 there have been new standards published jointly by the American Educational Research Association/American Psychology Association/National Council on Measurement in Education ("Standards for Educational and Psychological Testing") that provide guidance for test developers and test users. There are 24 different standards for validity.

Mrs. Doorey asked that Ms. Taylor walk through the process for one student taking the DSTP. Ms. Taylor explained that the student takes the test, tests booklets and answer documents are collected and shipped to Harcourt, they are opened and the answer document is scanned. The paper answer document then goes through an editing process where Harcourt looks for evidence of cheating, excessive eraser marks or any other alerts that it would be important for the schools to know about, e.g. writing about self destructive practices. The paper document gets filed at this point and all the activities related to that student are done electronically. The multiple choice questions are machine scanned and the items to be scored by a person are put in a queue.

The answer documents have all been bar coded so the machine scored portion of the test is matched up with the hand scored items. The points are tallied and then the raw scores go into the system where quality control and the equating study begins. It takes about three weeks to hand and machine score approximately 85,000 assessments (includes reading, writing and mathematics). The equating takes about 5 days and is done by a computer program but reviewed by three psychometricians, one from Harcourt and one from an independent firm. The third person was added two years ago for added "insurance."

Quality control includes things as simple as looking to make sure the answer key is the correct one, to analyzing if a question is missed by a large number of students. The students are given the benefit of the doubt if there appears there was a misunderstanding of the question and it was missed by a large number of students. This is done for both Delaware and SAT10 items. After equating and quality control, the scale scores then are related to a performance level. There is vertical articulation from grade to grade. We cannot determine grade level equivalence from our test (i.e., a student with a scale score of 500 and performance level higher than meets the standard in third grade could not be determined to be equivalent to a student in a higher grade where 500 meets the standard.)

A comment was made that there was a study by the University of Maryland that indicated 23% of students are misclassified. Ms. Taylor said that we have done our own studies and did not have that same conclusion. Ms. Taylor said the University of Maryland did not have the data to do this type of study accurately.

There was discussion related to the instructional needs comments. The instructional needs comments for the DSTP are not triggered by a performance level or scale score but by clusters of items missed. It was mentioned that the mathematics instructional needs comments were the most valuable and writing the least valuable. Unlike for reading and mathematics, the writing instructional needs comments are based on the student's total score.

A question was raised about the computer based assessments and validity and reliability. Ms. Taylor said the same quality control needs to be done. If it is for high stakes then more quality control needs to be done than if it is just to inform.

Discussion ensued about what happens with the scores. The scores are released from Harcourt to DOE near the end of May where the data go through more quality checks before being released to the districts at the end of May. The score reports usually get to parents by mid-summer. Printed copies of the score reports are shipped to the districts usually the second week of June. Those parents having kids that are required to go to summer school are notified by the school as soon as the data are released from DOE at the end of May. Some districts contract with Harcourt to mail the parent reports directly to the student's home, while others send them out themselves. Harcourt will send the score reports to parents, but there is a cost – this is a district decision. The actual writing images are available to the schools in mid-August. Different school districts have different policies related to teacher/administrator access to the password protected side of the DSTP online reports. Additional items available in the DSTP online reports include an electronic Individual Improvement Plan (IIP) template. Some districts and schools use these templates while others do not; they design their own.

David Sokola asked a question regarding the fairness of assessment for students with disabilities. Ms. Taylor and Mrs. Woodruff explained that there are three special populations: students with IEPs (Individualized Education Programs); 504 Plans which are for students with a major life function limitation; and LEP (Limited English Proficient) students.

Mrs. Woodruff explained there are many accommodations available for the DSTP. The goal is to ensure the student is given the assessment in the same manner accommodations are provided in the student's instructional program, for example, a student who cannot write and has a scribe would get a scribe for the assessment. A student who uses a calculator in class would get the same accommodation on the assessment. All of our accommodations go through our Technical Advisory Committee (TAC). There is also an allowance for students to be excused from the assessment on a case-by-case basis because of special circumstances including physical, emotional, medical or psychological issues.

Mr. Sokola said that the issue of assessment for special education students who are not severely cognitively disabled or high functioning, often referred to as the "gap" students is one he hears from his constituents. Mrs. Woodruff said that we are not in a position to develop an assessment for these students at this time, but that Delaware and other

states want to do this. Previously we could test students “out of level”, that is, a student in the 5th grade but instructionally and performing at 3rd grade was given the 3rd grade assessment. NCLB does not allow for this. Mrs. Woodruff expressed that this is something she and her colleagues from around the country would like to see changed with the reauthorization of ESEA in 2007.

The next item on the agenda was to talk with Dr. Steve Dunbar from the University of Iowa. He has been on our TAC since the beginning and is currently the Chairperson. The first question from the members asked his opinion of how Delaware compares with other states on the technical aspects of our assessment system. He said that related to reliability which is quantifiable, the state compares quite favorably against other states. He said Delaware has defined the open ended questions well and that allows for better scoring. In validity, for NCLB the alignment to the standards is the key ingredient. He said that this is hard to describe quantitatively and a lot of state have “force fit” their assessments which may cause some slippage between the standards and questions on the assessment. Dr. Dunbar thinks Delaware has an enviable model in terms of an assessment that comes out of a standards-based model. He noted that Delaware has also done some good work over the past year in looking at alignment across the grade levels.

David Sokola asked about the reliability of constructed response questions compared to adaptive computer based assessment. Dr. Dunbar indicated that if a test is truly adaptive, that is, trying to match the questions to measured ability, the thought is that you can get a good measure of the child’s ability with fewer test items. The presumed advantage of computer adaptive assessment is that you can give a shorter test and get the same reliability. He said this is dependent on the characteristics of the item pool, e.g., number of items in the item bank as well as the nature of the content/concepts being measured. He said if the item bank is shallow or heterogeneous then the advantages are not there. Dr. Dunbar indicated that for NCLB purposes where alignment is critical, you need to cover the content domain completely; it is tricky to do this with a computer adaptive framework.

Dorothy Shelton asked about the “off-grade” reliabilities. Dr. Dunbar indicated that the TAC members review the technical data every year and he has not noticed any dramatic differences between grades.

Dr. Dunbar said it was critical for groups to review tests and make recommendations periodically, like what this task force is doing. A comment was made that we want to move forward, and that it is critical we maintain the strength we have, make it better and be smart about any changes.

The next item on the agenda was the discussion of the costs of the DSTP per the documents handed out to the members (see attached). Based on the information, the Spring and Fall administration of reading, writing and math for approximately 85,000 students was around \$25.00 per content area/per student. For science and social studies and the approximately 34,000 students who annually take the assessment, the cost was about the same at \$24.00 per content area/per student. The retest for the diploma and summer school are not additional costs for development because what is given to the students is a former intact version of the test.

A question was asked as to how many new items are developed and field tested each year. Ms. Taylor said that there are 60 items per content area field tested in science, social studies, reading and mathematics. Twice as many items are developed but this number is honed down through the content and bias/sensitivity reviews. The writing prompts are written and benchmarked by Delaware educators.

Ms. Taylor also informed the task force that there was a large vertical and horizontal technical study not included in the costs. This cost approximately \$350,000. We also reviewed the performance level cut scores. Good policy suggests this be done every 5-7 years. Ms. Taylor said that all of this evidence has been sent to the US DOE for the mandatory peer review process.

A question was asked about the “item samplers.” Ms. Taylor said that sample items and samples of student work have been added to the “item sampler” portion of the DSTP online site.

A question was asked about any federal money we receive for developing the state assessment. Ms. Taylor said that we have received federal money for the last three years that is to be used to offset the cost of item development and the technology infrastructure.

The cost of a change to our assessment can not be determined until we know the changes. Costs will vary depending on the structure of the program and potential changes. Mrs. Woodruff said that the next meeting should focus on finalizing the purpose statement and a one page document that becomes the communication document describing the work and direction of the task force.

The next meeting is scheduled for March 15th in the Cabinet Room in the Townsend Building at 8:30a.m.

HJR 4 with SA 2 – Assessment Task Force
Meeting Notes - REVISED
March 15, 2006

Location: Cabinet Room, Delaware Department of Education

Time: 8:30 a.m.

Attendees: Jean Allen, Vicki Cairns, Cindy DiPinto, Nancy Doorey, Emily Falcon, Susan Haberstroh, Tony Marchio, Nicole Quinn, Noel Rodriguez, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Robin Taylor, Nancy Wagner, Valerie Woodruff

Members of Public: Charlene Tucker

Secretary Woodruff called the meeting to order. She distributed a letter from the US DOE stating the results of Delaware's peer review. Delaware's system received "Full Approval with Recommendation." The one recommendation related to development of performance level descriptors for science which the Department already had planned to complete in summer '06. She noted the involvement over the years of many educators and stakeholders in numerous DSTP-related meetings. She went on to note that many states that submitted their system for review have not shown evidence that their assessment system aligns with their State's standards.

To date, the task force has spent a great deal of time providing opportunities for experts to answer questions from the members. Mrs. Woodruff suggested we need to move toward detailing what it is we want. Nancy Wagner offered to extend the date of the task force reporting date, DOE is drafting the legislation.

Secretary Woodruff provided copies of two statements of purpose, one she drafted, and the other by Mrs. Doorey, Bruce Harter and DSEA. These documents were also sent via email prior to the Task Force meeting. Secretary Woodruff pointed out that the statement of purpose she created reflects the words the task force members provided in an earlier task force meeting exercise. Mrs. Woodruff reiterated that the task force be focused on assessment and not other issues, e.g., infrastructure. Mrs. Doorey noted that she and the other co-authors of the alternate version also worked directly from the task force's list of desired attributes, and then attempted to elaborate such that the meaning would be clear to non-educators.

Secretary Woodruff started the discussion by emphasizing the importance of the need for the technical underpinnings so we know we are doing the best job we can to use the assessments appropriately. Cindy DiPinto confirmed that if we do more than one assessment for our system then they all need to be aligned and they all must have the technical work done. Secretary Woodruff agreed stating that if we use assessments to make decisions, the technical underpinnings must be present. Compliance is not the issue.

David Sechler raised a concern that validity will be more complex because there are more tests. Secretary Woodruff noted that any formative assessment that is part of the system must also link to the standards just as the summative needs to.

Nancy Doorey suggested that the State continue to use multiple item types in its future assessment system but that it tie as little as required to NVCL in order to reallocate more funds to better help teachers and administrators improve student learning. She handed out the North Carolina feedback letter from their Peer Review – their approval is still pending. North Carolina has submitted their core assessment for approval for NCLB. In addition, the State uses seven other assessments for state/internal accountability. She brought up the question of allocation and emphasis on

primary and secondary purposes of assessments and their use. Secretary Woodruff indicated that the DOE does not strive to do average work therefore, does not want to meet only the minimum requirements of NCLB.

Robin Taylor emphasized that NCLB is important but has not been a driver for the current Delaware assessment system in place. Delaware developed its system prior to NCLB. The improvements since then have been beneficial and needed to be done. These improvements certainly helped with NCLB compliance. For example, the DSTP Technical Advisory Committee (TAC) gave advice to Delaware for the peer review process in regard to technical documentation needed. Other states had a very hard time with this. Wendy Roberts reminded the Task Force that the ongoing discussions and meetings with the test development committees focus on the linkages to DE content standards, not on NCLB.

Secretary Woodruff emphasized that NCLB has been helpful because:

- Delaware needed clearer grade level expectations. This guided the recommended statewide curriculum.
- Delaware needed clearer performance level descriptors. Recently, we moved from SAT9 to SAT10 which gave us the opportunity to look at and revisit the DSTP performance levels for all grades (cut scores).

She also stated that we would have done these things regardless of NCLB.

Nance Wagner raised concerns. She does not want the purpose of the Task Force to be focused on just one assessment and one prescription with a narrow direction. She wants the recommendation to be open and guiding. Secretary Woodruff responded by saying we all agree that a single assessment does not work. We need a system and it is important that the task force think about what that system could look like.

Mrs. Woodruff suggested that we need -

1. some kind of summative assessment that can be used for a variety of purposes.
2. more information on a regular basis for teachers, students, etc. [She noted that there are a lot of formative assessments out now that weren't in place a few years ago.]
3. possibly end-of-course assessments could also fit into the system.

We need to figure out what purpose all of these serve and if there is anything else to consider. Also, the task force needs to decide the entity responsible for the various parts of the system, that is, state, district, school, and classroom. We need to think about how classroom assessments fit in and about what we can do to help improve our teachers assessment literacy.

Nancy Wagner feels as if she is being pushed toward the pilot plan and is concerned with the dollars that would have to be allocated toward that plan. She noted that we have to have a system we can afford.

Nancy Doorey felt the following concept is being left out of the minutes and the planning: "Accurately measure individual student growth on a vertical scale over time towards the standards." Ms. Doorey feels this would give much more powerful information to teachers, administrators, etc. Robin Taylor stated that a vertical scale is a test characteristic, not a test purpose. There are tests that will measure vertical scale. She noted that this does increase the amount of technical work that must be done and also the cost. Nancy Wagner stated that the purpose should not be test specific.

The group decided that whatever statement of purpose is used, it should include the statement: "Allow for the establishment of individual student goals and accurately measure each child's individual growth over time based on the standards."

Secretary Woodruff emphasized that the two statements of purpose need to be blended. Tony Marchio reminded the group that the format should move away from technical language and be targeted to an audience of parents. A

question was raised regarding who the statement was for. It was decided by the group that it is used to guide the Task Force's work and it goes to the governor and the legislature. Secretary Woodruff restated that the statement is for the Task Force and does need to be in plain language. However, for now, she chose words used by the Task Force.

Secretary Woodruff pointed out that the "Administrator and Teacher Coaches" segment of the Doorey/Harter/DSEA drafted statement deals more with infrastructure rather than the assessment system that provides the infrastructure. Bullet number two in the DOE drafted statement is a better statement because DOE chooses the reporting it wants from the testing vendor. The testing vendor simply provides the data based on our design requirements. If we want more frequent reporting then we need to talk about components. However, now we are in the assessment aspect of our discussion. Secretary Woodruff noted that quality checks at the end to ensure accuracy are still needed; there is a concern not to label schools/districts inappropriately. Secretary Woodruff stated that summative assessments give us information in a certain timeframe but if we talk about other components of the system then we begin to think about formative tests.

Nancy Doorey agreed with the shortcomings in the "Administrator and Teacher Coaches" segment of the draft she provided. However, she stated that she believes DOE spends a great deal of money to give information on decision-making. If we are to think about next generation systems rather than districts doing the reporting, she believes we need a more powerful web-based tool. Nancy Wagner stated that this Task Force is not designed to make that decision. Secretary Woodruff stated that the steps of this Task Force are to determine the purpose of the assessment and what components we want. She reminded the task force that the components do not necessarily have anything to do with the infrastructure and that once the components are established then we could think about the infrastructure and reporting system.

Noel Rodriguez reminded the Task Force that from a parent and administrative perspective the community needs to know what we are trying to assess in a simple way. We cannot use jargon with parents.

Nancy Wagner wants to be sure what we do is what we need; do we really need all the data we are discussing. In the real world, will people really access and use this data? Secretary Woodruff agreed and stated that we need to figure out why we want the data and for what purpose. David Sechler does a great deal of data manipulation and stated there is a lot of data in the DSTP online reports that you can use to inform instruction if you become familiar with it. However, he noted that it does take time. He had to develop templates and spends about two weeks in the summer summarizing the data. He gives the data to his staff in these templates and the staff analyzes it. This process brings up interesting things. He finds the public side of the DSTP Online Reports to be useful for this purpose. This is the professional development piece that comes after this Task Force.

As a side note, Secretary Woodruff stated that the DSTP Online Reports has been around for a while and right now the entire system is being refined to make it more useful and efficient. She also mentioned the work of the University of Delaware's Research and Development Center on the "correlates of achievement" and the Vision 2015 group that is currently doing a review of the public education system.

Tony Marchio mentioned that the data is tremendous and support from the state is exceptional but that an area for improvement would be at the teacher level. Robin Taylor noted there are two parts to professional development that will improve instruction and improve the assessments teachers give. These are assessment literacy and use of data.

David Sechler stated that the system is important because no one component gives us what we want. The test may not be designed for NCLB but when administrators at the school level see their NCLB status, it affects them. He suggested that what we do and what our status is has a strong connection. Secretary Woodruff responded by saying

that although there is a focus on NCLB, the State did have an accountability system in place. The first school rating was done in 2001. We did have to change the system when NCLB was implemented.

Dorothy Shelton provided comments that she feels students should be compared nationally as well as across the district and the State. In addition, there needs to be a commitment for parents to get good information so that they are able to support their children. She feels that the system must include multiple things (e.g., a summative assessment and formative assessments) whether it is organized with some parts being the State responsibility and some parts the district responsibility.

Secretary Woodruff posed several questions to the task force. She asked, in the system of assessments, what information should the assessments provide? We know we need a summative aspect. We agree that formative assessment is important to regularly inform students and teachers about student progress toward standards. What role does an end-of-course assessment play? What do they do for us and at what level should they be administered? The task force agreed that end of course exams should be administered at the high school level and some said they should be administered at the middle school level, but only for courses that lead to a high school credit. A conversation began around the concern about an overabundance of testing at one particular time of the year and if this extra layer of testing is in fact valuable. Is an end-of-course exam not a summative assessment? The answer was that an end-of-course exam is specific to the subject while a summative exam applies more to entire grade learning.

Nancy Wagner stated that final exams, or end-of-course assessments, are beneficial because they are a post-test, they give data on teacher performance, and they force kids to be engaged. Secretary Woodruff stated that end-of-course exams will not fulfill the entire summative piece because they are too narrow. Dorothy Shelton added that there are a variety of forms of subjects so it will not be easy to design. Robin Taylor stated that for the purposes of accountability we could add up the pieces to make it summative on a broader level as long as the tests still measure the State's standards. For example, Algebra I + Geometry = Accountability Score. The score on the end-of-course assessment could be a part of the final grade.

Noel Rodriguez stated that from an administrator's perspective it is important to assess after every grade level but as a parent, what are the consequences if a student fails? And what is the school going to do about it? These repercussions need to be consistent throughout the State to avoid controversy. Secretary Woodruff stated that if we have consistency then the State begins to dictate what a school must do in terms of content. The recommended statewide curriculum being developed is a framework.

Nancy Doorey stated that nine states use end-of-course assessments for NCLB and that some of these states use the end-of-course assessments as part of their final grade. She suggests that end-of-course assessments be used as our high school summative assessment. It is far more motivating to kids if an end-of-course exam is our summative assessment tool in high school. The current summative assessment, the DSTP, is given when students are in tenth grade and may not reflect the material a student has covered. The benefit of an end-of-course assessment is that students take an end-of-course exam when they take the course. In middle school, if you get a high school credit for a class then you can take the end-of-course exam and bank it for high school.

Vicki Cairns raised a concern that the State would be specifically prescribing what would have to be taught in a course. She asked, what will happen to a student if it takes two years to pass the end-of-course exam. Secretary Woodruff reminded her that time is the variable in this equation; the student could be given more time and then take the test again. The issue becomes if the end-of-course exam is more restrictive for a teacher more than the fact that teachers should be teaching to standard.

David Sokola suggested that districts have a “menu” of assessments that a district could choose from. This started a conversation in the task force regarding whether end-of-course exams should be in addition to or substitution for the DSTP. If the end-of-course exam is approved by DOE then districts could use it as a substitution or the district could decide to have multiple tests. Dorothy Shelton asked if the State could delegate some choice to the local level depending on the assessments available. Mrs. Woodruff stated that we have to keep in mind that if an assessment is used for accountability purposes there has to be technical underpinnings and security.

Nancy Wagner added that an end-of-course exam should be a display of the students’ understanding the learning concepts taught by the class, not memorizing the main character in a specific play discussed at the beginning of the year. End-of-course assessments should be taken at the same time across the State for security. In addition, end-of-course exams will allow for professional development and student achievement.

Dave Sechler stated that he did not see a big difference between the DSTP and end-of-course assessments at the middle school level.

Nancy Wagner stated that if scoring is a problem we might consider collaborative scoring. Dave Sechler noted that when this method was evaluated, a broad range of scores were given by scorers on the same test. This method has shortcomings.

Cindy DiPinto expressed a concern that some students might not understand course content completely until they take more advanced courses the next year, with an example of not understanding Algebra until after Geometry was taken. She did not want to see the system be too rigid.

Secretary Woodruff reiterated that how the assessments are used still continues to be the main question. How will end-of-course assessment count toward learning? Will it be a percentage of the final grade or will the assessment be blended into the regular final test? Dorothy Shelton noted that if the end-of-course assessment were to be part of the final grade for the course it would have to be scored with a quick turn-around; it would probably be shorter than the DSTP, perhaps administered on the computer, etc. A comment was made that if we are going to have scoring information back quickly then we cannot use the DSTP. Also, if technical underpinnings and quality are present, scores would not be instant. They could take a day to a week.

Tony Marchio stated the further the teacher is away from the assessment the less meaningful it is for them. He said ideally if teachers score the assessment, it will be more meaningful.

Nancy Wagner stated that she strongly prefers end-of-course assessments given at the end of the school year/course at the high school level, and that they need to be a replacement for the DSTP and not in addition to the DSTP.

Nancy Doorey proposed that the end-of-course assessment in high school be two-fold with the essay given in the spring and the constructed responses and multiple-choice given at the end of the year/course. The multiple choice and writing questions could represent the NCLB portion and all else could represent the State’s portion. Robin Taylor added that the first part of the test could be administered in April for NCLB and the rest of the test could be given at the end of the school year. The two could then be tied together to get the final score. The open-ended portion could be scored locally. Mrs. Doorey suggested if only the multiple choice items are tied to NCLB for reading and math, that you could use retests from summer school to recalculate accountability, according to her conversation with Kerri Briggs at US DOE. At this time, we cannot recalculate scores using summer school scores because this has not been decided for NCLB. Since the quality of multiple choice questions has improved over the years, it is possible to construct a test that measures a student’s reasoning skills. There will always be a portion of luck.

A concern was raised about value. If we create a system where one portion is for NCLB and other portion is for the State will we run into a situation where students discover that and begin to not value the NCLB portion? The answer is “no” as long as the entire test is used for accountability purposes.

Mrs. Doorey noted that there has got to be an ability to compare performance across state lines. She also asked if we can do some types of items, such as the student constructed response and essay items, at lower cost if they are not used for high stakes.

Mrs. Woodruff stated that these other state assessments can be used with the new diploma legislation and used as a piece in how students can get a diploma. This allows students that do not test well to show their ability. This concept always has a lot of public support. In the end, the student could portfolio their ability. Mrs. Woodruff also noted the need for consistency across districts. Cindy DiPinto added that consistency applied to criteria and grading.

It was reiterated that we need to do what is best for kids.

Nancy Doorey suggested that in the interest of time the task force break extend an invitation to individual members or subgroups to bring forward rough draft proposals for the system of assessment, for full group discussion (Dr. Harter had previously made this request). Secretary Woodruff was cautious about this idea because she prefers everyone talk around the same table. However, she said that if members want to prepare draft proposals they should flow them back to her office by Thursday, March 30. This could provide time for the department to present the information in a consistent manner at the next meeting.

Next meeting dates to be held in Cabinet Room:

April 3, 2006 at 8:30 a.m.

April 12, 2006 at 8:30 a.m.

HJR 4 with SA 2 – Assessment Task Force

Meeting Notes

April 3, 2006

Location: Cabinet Room, Delaware Department of Education

Time: 8:30 a.m.

Attendees: Vicki Cairns, Cindy DiPinto, Nancy Doorey, Susan Haberstroh, Nicole Quinn, Noel Rodriguez, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Robin Taylor, Valerie Woodruff, Ellen Forte, Bruce Harter, Yvonne Johnson

Members of Public: Cecelia Le, Charlene Tucker

Supporting Documents: What is the Evidence that Use of Multiple Choice State Tests Negatively Impacts the Teaching of Complex Reasoning and Higher Order Thinking Skills? (Appendix F), Delaware Statewide Student Assessment System Presentation (Appendix G), and “Straw Man” Proposal for the Next-Generation State Assessment System Presentation (Appendix H)

Secretary Woodruff called the meeting to order. She distributed meeting notes from the February 27, March 3, and March 15 meetings and a document outlining cost for the DSTP. She informed the task force that the reporting date is extended until May 31st. Mrs. Woodruff told the task force that hoped that the day’s discussion could begin to lay out some recommendations that could then be translated to a Request for Proposal. She introduced Dr. Ellen Forte who will help lead the day’s discussion.

Dr. Forte provided a PowerPoint with handouts. (See handout for detailed information regarding her presentation.) The PowerPoint outlined what she felt had been agreed upon by the task force to date. Specifically that it is a system of assessments that is needed, and that one assessment cannot do all of what is hoped to be accomplished. She reiterated the statement of purpose, discussed support and stakes, validity burden, shared responsibilities, features, and gave an option to a system of assessments. She discussed the costs to some extent and said that the costs come down to the validity burden. In the case where an assessment is used for high stakes decisions, i.e. graduation, promotion, school, district accountability, validity of the assessment is critical. She talked about the possible role of the state versus the district and school roles in a revised assessment system.

There was discussion related to end of course assessments that measure the standards, discussion related to the delivery format, that is, computer versus paper and pencil. The mode of the test delivery must be fair and equitable. Dr. Forte stated that there is a body of research that discusses this topic. Secretary Woodruff was concerned because while the DSTP has accommodations other exams do not. Nancy Doorey handed a literature review on computer based exams and accommodations. One concern is the ability of students to “go back” or “skip over” questions in a computer based format. Dr. Forte also stated the young students have a hard time using the computerized tests because transitioning is a skill that needs to be learned.

Cindy DiPinto raised a concern regarding the chance that the student performed better on a paper test then on a computerized test, but there is no diagnosed disability that required the student to take the paper exam. The group stated that teachers are always trying to decide which mode is better for children.

Dorothy Shelton mentioned that if we use computerized exams then we can give the exams in multiple languages. Dr. Forte agreed, stating that the directions can be read in other languages and studies, but you must be cautious about translating assessment because it may change the construct of the assessment which affects the validity.

Noel Rodriguez stated there are a lot of students with diagnosed disabilities but parents do not want to label them so teachers are not aware. DOE must inform parents that the downfall to this is that their child will not receive accommodations. The group stated that this problem will probably always exist.

Bruce Harter and Nancy Doorey presented a “Straw Man” proposal (see handout for detailed information.) This proposal was developed with input from Nancy Doorey, Bruce Harter, Vicki Cairns, Martha Manning, and Edie Corbin. The proposal discussed an outline for a new assessment system which they believe could reduce NCLB costs, would test full school year, and provide information that could better inform teachers, parents, and schools. One of the themes in the proposal was to shift cost from the assessment to improving teacher quality. The group noted that the model was based on showing individual student growth over time and modeled to some extent on North Carolina’s assessment system. There was discussion related to decoupling the student constructed response items from the multiple choice items.

There was some discussion related to North Carolina’s assessment system as compared to their NAEP scores. Dave Sechler commented that NAEP is not necessarily valuable to him because it does not provide individual student data. Secretary Woodruff commented that NAEP only gives statewide data. In Delaware, the sample used in NAEP is very large and that NAEP tests almost all of Delaware students so it is not necessarily accurate to compare Delaware students to students in other states. The DSTP is NAEP-like regarding how the items are constructed.

Dave Sechler mentioned that his school’s writing scores had been better before NCLB. He believes that NCLB and the focus more on reading and math and less on writing decreased the value of writing in the testing process. In the end, a trade off occurred. He now has students that are doing well in reading but not in writing because he believes the teachers changed their focus.

Dorothy Shelton stated that the group agrees that the criticism of the DSTP is the time issue. This proposal solves some of that. She has seen a difference in instruction because of the DSTP especially in writing instruction and she believes that Delaware cannot afford to lose this skill. She stated that writing should be included in some way in the accountability system. Some of the assessment must be text-based. She also mentioned that the NCLB portion could be made simpler in that not everything we test has to go into NCLB.

Secretary Woodruff shared some concerns about the proposal. The proposal is assuming cost will be lower than the DSTP. However, this is not necessarily true and assumptions cannot be made until the RFP is out. Also, the district in the assessment pilot (NWEA) had to invest \$200,000 in 10 schools to get the computer aspect of the assessment up and running. Delaware needs a case study about the technological investment needed to make the switch. It was noted that several years ago, the Christina School district did an online pilot of the DSTP. The pilot was not very successful because of technology at the time.

Dr. Forte reiterated that alignment of the assessment to the standards is the ultimate criterion. She suggested the task force be cautious about using multiple-choice for high stakes exams because of the technical underpinning.

Questions about cost arose and Secretary Woodruff explained that the costs from an assessment come from test development, printing, scoring, benchmarking, and other costs but cautioned about getting sidetracked by cost. She advised the task force to focus on the system and then the costs can be addressed. The vendors may be helpful in providing alternatives that are more cost efficient that the task force may not be aware of.

Cindy DiPinto asked if there was student accountability built into the proposal. The group agreed that there was. There was also discussion about extending accountability into the 12th grade. Bruce Harter replied that if a student passes the exam in 10th grade then he/she will not have to deal with the assessment (state) but could then focus on other things like the SAT, AP or other types of activities.

Yvonne Johnson reminded the group that the SAT now includes writing. As a parent representative the most important thing is that the test is fair. Cost is not what a parent focuses on. She also supports computerized tests close to the end of the year because there is a lot of dead time that needs to be filled during that time.

David Sokola, looking at the surveys listed in the “Straw Man” proposal, stated that the group cannot ignore the positive comments from the teachers and administrators on the computer adaptive assessment being piloting in several districts and charter schools.

Secretary Woodruff in wrapping up the meeting summarized the main points of agreement within the group:

1. Formative assessments should be used throughout the year to inform instruction and give feedback.
2. There is an interest that assessments should be computer based understanding that the technological infrastructure must be provided to all schools.
3. The group values writing and constructed responses and the test should be more than just multiple choice.
4. The group needs to look at end-of-course exams at the high school level. These should be given when the student has finished the requirements. They should be standards based not calendar based.
5. Teachers must be engaged in the assessment process, particularly in the scoring piece. There is a concern that since Delaware is such a small state this will cause issues with perception and credibility.
6. There is value in getting as much as can be done close to the end of the year.
7. Professional accountability should be tied to some part of the process.

She also mentions that what will count for NCLB will come later in the discussions, after the assessment system is determined.

Also mentioned was that Delaware was one of eight states able to move forward in the process toward approval of a pilot growth model for AYP purposes.

Next meeting dates to be held in Cabinet Room:

April 12, 2006 at 8:30 a.m.

April 21, 2006 at 8:30 a.m.

May 17, 2006 at 8:30 a.m.

HJR 4 with SA 2 – Assessment Task Force

Meeting Notes

April 12, 2006

Location: Cabinet Room, Delaware Department of Education

Time: 8:30 a.m.

Attendees: Vicki Cairns, Cindy DiPinto, Nancy Doorey, Susan Haberstroh, Nicole Quinn, Noel Rodriguez, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Valerie Woodruff, Bruce Harter, Martha Manning, Nancy Wagner, Jean Allen

Public: Tom Ealy, Harcourt

Secretary Woodruff called the meeting to order and distributed a document that she drafted based on the previous meeting discussions. She explained that this could be used as a framework for the discussion and hoped that the day's meeting would get the task force to some decision points so that recommendations could start to be formulated. She asked the group to review the first three paragraphs.

Secretary Woodruff led the conversation by reviewing the two major components that the group agreed upon. The first was the need for informing instruction. Mrs. Woodruff suggested that the state could provide the infrastructure for locally administered formative assessments but wanted comment from the task force on which content areas these assessments should be in. Mrs. Woodruff reminded the task force that formative assessments are currently embedded in the science kits and units and social studies is being developed in the same manner.

Nancy Wagner requested reading, writing, and math assessments, especially in light of science and social studies embedded assessments. She said that if teachers can get information on an ongoing basis in reading, math, and writing, then he or she can get an idea of how the student is doing in other areas.

Dorothy Shelton clarified that the formative assessment is to inform instruction and she voiced the concern that social studies and science curricula are not aligned. She said that we need to know how our kids are doing in that social studies, not just reading, writing, and math.

Cindy DiPinto pointed out that while the Delaware Department of Education is continuing to develop a statewide recommended curriculum with embedded formative assessments, there is no guarantee that the districts and school will use these. The group agreed that the districts are using them even though they are not required. Cindy was concerned with the fairness of the assessments. She wants to be sure that students in all districts are being assessed equally. The group felt as if this was not an issue because there is no accountability attached to those formative assessments. They are embedded simply to assist teachers with instruction.

Nancy Doorey felt as if we need a benchmark, a trajectory toward standards. Social studies should be embedded, as is science and flexible and district driven. Reading, writing, and math should be benchmarked assessments.

Dave Sechler was concerned with formative assessments and when they are given. Formative assessments need to be embedded into the curriculum, as is with science and social studies. He was concerned that not all content areas are sequential and math, although some believe this to be sequential may not be to all students. He suggested we need to be careful when we administer assessments.

Vicky Carnes asked who develops formative assessments. Secretary Woodruff pointed out that the science model has worked well however we need to get social studies there too. In more districts we are seeing more department assessment or school based assessments. The statewide recommended curriculum will provide a more cohesive look at the district level.

Bruce Harter discussed California as having different content area teachers teaching the various types of writing. This assists with cross curricular teaching and doesn't leave writing to just the English teachers. Secretary Woodruff stated that we could do something like this and have specific types of prompts embedded in the assessments; however, there needs to be ownership at the school and district level.

David Sokola asked if there was training in the California model. Others noted that training and coordination is needed. Bruce Harter said that training, coordination, and collaboration are certainly part of what was done in California. Secretary Woodruff suggested that one of the recommendations from the task force will be to encourage more collaboration. Mrs. Woodruff commented that she had heard that other legislators were concerned that there continue to be local control.

Noel Rodriguez pointed out that it is very important that there is ownership across the entire curriculum. Students need to be engaged in science and social studies and there needs to be fairness. Nancy Doorey stated that if infrastructure is in place that will prompt a recommended curriculum, electronic, e-school, blind scoring anchor. We just need to provide technology or infrastructure. Secretary Woodruff was not sure if technology is all that is needed but rather anchor papers created as a team at the school/district level. Conversation ensued around benchmarking, anchor papers, and teachers in the schools.

Bruce Harter was concerned about the logistics around technology and group scoring.

Dave Sechler stated that we need to develop staff to get the students beyond what is taught. That takes time. Secretary Woodruff believed that we could affect professional development that occurs in schools with models that we develop.

The following points had consensus from the group:

- Science and social students will be within the curriculum units. The state will provide models however districts and schools will be responsible for the assessments.
- Writing needs to be reinforced in all content areas in all grade levels. Again, schools will be responsible for the assessment and the state will provide models.

Secretary Woodruff stated that reading and math, as discussed in past meetings, would be "on demand" assessments. Districts will control when the assessments are given but the group seems to be saying that the assessments should be standardized with a bank of items.

Dorothy Shelton stated that she would like adaptive assessments so teachers can use the information. Dave Sechler stated that he liked NWEA because it gave him a cluster of comments about a student.

Nancy Doorey mentioned Oregon's assessment which could be approved in May. Oregon gives "on demand" assessments that combine paper and pencil tests with computer adaptive. It will be interesting to see if this type of system gets approved. She also stated that it is important that we assess at the beginning of the school year, the end of the school year, and "on demand" in between. This will allow us to see growth during the school year.

The following points had consensus from the group:

1. We need to be sure that formative assessments are adaptable to the student's level so that we know where they are in skills and knowledge. They should also tell us what needs to be addressed so we can get the student to the standard, if needed.
2. The assessment needs to meet the requirement of testing students by the end of the year at the enrolled grade for accountability purposes. Testing students out of their grade level does not meet NCLB requirements.
3. Writing should be assessed in some manner every year for each student.
4. The state ELA standards include both reading and writing
5. Math is needed annually for each student.
6. Each student does not need to be tested annually in science and social studies.

A question still was on the table regarding how the formative fit into accountability.

Dorothy Shelton stated that the test coordinators say they want science and social studies to be a simpler assessment. Currently it is too long and they would like to shorten or eliminate the text-based responses in some or all of the assessments. The test coordinators prefer the assessment to be administered at the end of grades 3, 5, 8, and 11. Dave Sechler asked that the tests be administered in the grade which it is taught.

Secretary Woodruff mentioned that it has been discussed in the past to move the science and social studies assessments to the end of 11th grade instead of 10th. This could create an issue with the end-of-course assessments. If we choose to do a single assessment, it could be moved to the end of 11th grade which gives teachers and students more time to prepare. However, because it is tied to student accountability this timeline will not give students a lot of time to retake the assessment, if needed, for graduation. There was also mention that the students take the end-of-course assessment after the class. There would then be more tests that would count toward student accountability. There could be a combination of assessments for student accountability.

David Sokola mentioned that if NCLB does not care which grade the tests are taken, then the students should just take them when they are ready.

Nancy Doorey likes the high school end-of-course model because there is a spectrum of courses. She mentioned the North Carolina (NC) model because writing, reading, and math are applied to NCLB and the others are applied to the attainment of the student's diploma. In addition, NC is in the process of having a graduation project as a requirement. Secretary Woodruff added that other states tie the end-of-course assessments to the course grade, much like a final exam. This ensures that the student is invested in the process.

Dave Sechler was concerned that end-of-course assessments create a hierarchy of courses. Also, he asked that if we change the assessment to the 11th grade does the assessment become more difficult or does it remain at the 10th grade level? Secretary Woodruff responded that we can have expectations at the high school level that reflect our standards. The grade level piece is when the student takes the test. It just cannot be given in the 9th grade. Again, the advantage is the student in the 11th grade would be more advanced in their learning process when they take the assessment.

Secretary Woodruff reminded the group that we are changing the high school experience.

Nancy Doorey mentioned that in Wisconsin if a student got a certain number of high school credits he/she could then take classes at the community college. The group stated that Delaware does this as well.

The following point had consensus from the group:

- The group values the idea of end-of-course assessments in high school.

Nancy Doorey asked if the Task Force could get feedback from the graduation requirement committee. Jean Allen stated that since the committee is trying to make high school more rigorous she believes that the end-of-course assessments would compliment their recommendations. She, of course, is not absolutely certain of their viewpoint. She did state that in their last meeting there was consensus that the districts should be given the freedom to have their own plans for making the senior year meaningful for their students with some oversight from the state. Secretary Woodruff agreed. You do not want to hold students to one requirement if another is a better suit for the student.

Dorothy Shelton stated that if we take the end-of-course assessment approach, what happens with participation and cells for AYP? Secretary Woodruff responded that we will work that out. If we want high school to improve then we must look at it in a different way then what is being regulated by NCLB.

Bruce Harter stated that if we want to have alignment with state standards and accountability this will allow us. We have to create more rigor rather than dummyming it down.

David Sokola stated that the NC end-of-course exam had a grade component to it. In Delaware, end-of-course exams are a good combination that would appease parent and business groups. It would eliminate the argument that, for example, Algebra I is more difficult in one school then another. Cindy DiPinto mentioned that it will also address the issue of mobility. However, end-of-course exams must be graded equally.

Nancy Doorey mentioned that NC has students from grades 3 to 8 take end-of-grade exams and high schools take end-of-course exams. Students take the assessment two weeks before the school year ends. The grades come back the same day and low performers are grouped together with a teacher to prepare for the next retake of the exam given five days later. Fifty percent of these students pass the retake and do not have to attend summer school.

Secretary Woodruff reminded the group to let the vendors tell us how to get where we want. We do not want to say we need it done a certain way. We can give preferences. She also is uncomfortable with having students retest five days later. These students need to learn over time rather than the content being drilled in their heads over a short period of time. However, Nancy Doorey disagreed. She stated that if you know where the student is failing, a teacher should focus on that particular area.

Nancy Doorey also pointed out that the second paragraph in the distributed document should add “minimize costs by working with other states.”

Secretary Woodruff brought up the subject of multiple-choice. The group had made a decision prior that it should not just focus on multiple-choice. The short answers can be scheduled earlier on in the year and the multiple-choice could be scheduled closer to the end of the year. She asked the group how we deal with valuing short answers and the timing issue.

Dorothy Shelton wanted to do more complex multiple choice and a few extended responses.

Secretary Woodruff stated that the stand alone prompt is valuable. She brought up the topic of doing multiple-choice toward the end of the year and doing the text-based out of the various content areas. For example, we could do technical writing off of a science prompt, another writing off of a different prompt, and so on. Therefore students are giving extended responses, but only a few early in the year. Dorothy Shelton agreed. The items could then be

released to the teachers before the end of the year. A question was posed by Nance Doorey as to what level it would count toward because we do not want to add to NCLB.

It was stated that students can retest on NCLB portions but not on the portions of the test that are not tied to NCLB. That portion will apply to state accountability the same as social studies does. As long as the test is tied to some type of accountability, like summer school or graduation, the student will value it.

The group was reminded that we care about students more than NCLB. Delaware has always put students before federal requirements. Assessments do not have to feed into NCLB but accountability does need to be tied to them.

Secretary Woodruff asked the group to think about how to connect extended responses. She also asked at what grades to we do end-of-course assessments (cannot do them 9th or 12th grade), what do we test, and how many do we provide. Also, should the end-of-course assessments include extended responses?

Dorothy Shelton stated if we link assessments to NCLB we are boxed in with no flexibility. The group agreed with this important point.

Nancy Doorey wanted to hear from the curriculum cadres regarding what design yields the most learning. Dorothy replied, stating that student constructed response is most helpful.

Next meeting dates to be held in Cabinet Room:
April 21, 2006 at 8:30 a.m.

HJR 4 with SA 2 – Assessment Task Force

Meeting Notes

April 21, 2006

Location: Cabinet Room, Delaware Department of Education

Time: 8:30 a.m.

Attendees: Jean Allen, Vicki Cairns, Cindy DiPinto, Nancy Doorey, Susan Haberstroh, Bruce Harter, Martha Manning, Noel Rodriguez, David Sechler, Dorothy Shelton, David Sokola, Robin Taylor, Valerie Woodruff

Public: Charlene Tucker

Guest speaker: Phoebe Winter

Supporting Documents: An Evaluation of the Alignment of Three Formative Assessments to Delaware Grade Level Expectations Presentation (Appendix I), Next-Generation State Assessment System (Appendix J), and Assessment System Components (Appendix K)

Secretary Woodruff called the meeting to order. She noted that on the day before she had sent out a document “An Evaluation of the Alignment of Three Formative Assessments to Delaware Grade Level Expectations in Reading and Mathematics for Grades 3, 5, 8, and 10.” This is a report that had been prepared by Phoebe Winter, Ph.D. She noted to the task force that they were the first to see this document.

As had been discussed in previous meetings, Mrs. Woodruff reiterated the importance of alignment to the Delaware content standards in any assessment we develop. Mrs. Woodruff started the conversation by stating that the state had contracted with Dr. Winter to do an initial alignment study of some of the formative assessment currently available. Six vendors were invited to submit information; three of the vendors participated. The others did not participate for various reasons including the inability to meet the deadline for submitting their items. A question was raised about doing an alignment study so early in the process since the task force was looking at formative assessments on a conceptual level. Mrs. Woodruff and Robin Taylor stated that this is just the first attempt to see what is available “off the shelf” and is not intended for any type of selection of a potential assessment vendor in the future. There are many assessment vendors and this was not an in depth study. The methodology is nationally recognized and was used in the alignment studies for the DSTP.

Dr. Winter is an independent consultant and does a great deal of work with the Council of Chief State School Officers (CCSSO). She provided a PowerPoint presentation. She said that now that most states have satisfied the requirements of NCLB, there has been more of a focus on formative assessments. She noted that Delaware was a leader in this area. She noted that assessment alignment consists of “the degree to which an assessment measures the content of the standards, at an appropriate level of complexity, with an appropriate degree of emphasis on the objectives (performance indicators) within the standard.”

Dr. Winter explained the process. The vendors were asked to provide test forms with the “most frequently exposed items in Delaware at each grade level.” The test forms were going to be analyzed to determine alignment to the Grade Level Expectations (GLEs). Mrs. Doorey wanted clarification that the vendors may have submitted items that were “most frequently exposed” but that may not be indicative of alignment to the GLEs because of students getting questions while not at grade level. One of the purposes of a computer adaptive formative assessment is to determine where the student is performing, which may or may not be at grade level. Dr. Winter stated that for purposes of this study, the analysis is against the GLEs at the select grades of 3, 5, 8 and 10 in reading and mathematics. There was

agreement that the language in the report is correct. There were also comments related to being clear in language and intent on the Request for Proposal.

The three companies who chose to participate included: *Measures of Academic Progress – Delaware Version* by the Northwest Evaluation Association; *Progress toward Standards* by Measured Progress; and *Stanford Learning First – Class Views* by Harcourt Assessment, Inc.

The methodology used in the alignment procedure was developed by Norman Webb. His procedure is widely recognized in the assessment and psychometrical community. Dr. Winter said that there was a great deal of time spent in training the 33 retired and active Delaware educators who participated in the alignment study.

The four characteristics of Webb's criteria for alignment include:

1. **Categorical Concurrence** – Indicates the degree to which the same or consistent categories of content appear in both the standards and assessments.
2. **Depth of Knowledge Consistency** – Indicates the degree to which there is consistency in the cognitive demands of state standards and the cognitive demands of the assessment items.
3. **Range of Knowledge Correspondence** – Indicates whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities.
4. **Balance of Representation** – Indicates the degree to which items are appropriately distributed among the objectives.

Dr. Winter explained that the educators worked independently and then were brought back for discussion and followed by a debriefing process.

The four criteria for evaluating alignment include:

1. **Categorical Concurrence** – Required at least 6 items per standard.
2. **Depth of Knowledge Consistency** – Requires at least 50% or above the DOK of the performance indicators.
3. **Range of Knowledge Correspondence** – Requires at least 50% of performance indicators measured by at least one item.
4. **Balance of Representation** – Requires an index value of at least 0.70 (1.0 would indicate perfect balance of representation).

Dr. Winter noted that Categorical Concurrence is the foundation for alignment.

The information from these types of studies is used by vendors to tailor and custom design tests for the particular state or client. This study provided a good concrete idea of what is available currently without customization. Dr. Winter said that each of the three tests could be modified to match Delaware content standards and grade level expectations but also said that constructed response items would also need to be added.

A question was asked whether all formative assessments are computer adaptive or computer based OR paper and pencil. Ms. Taylor said that most are computer based or computer adaptive but there could be instances of a bi-modal assessment.

Mrs. Doorey asked if Dr. Winter knew where states were in respect to their summative assessments and alignment. Dr. Winter did not have the information available. Ms. Taylor said that alignment of content to assessment has always been done, but that the Depth of Knowledge and Balance of Representation were two new pieces of alignment. Ms. Taylor said these studies are critical for test development.

Mrs. Doorey commented that the purpose of the formative assessment is to provide information to teachers as to where students are (i.e. grade level) and not only against GLEs. There needs to be the ability to place the student in the continuum of the standards across grade levels so that instruction could be tailored to the student. Mr. Sechler noted that it is important for him to know what skills and knowledge the student needs and not at which grade level the student is performing. Mrs. Shelton and others noted that we need to know about the students that are above grade level also. Mrs. DiPinto commented that we need to be clear to the public that kids are not expected to know the answer to all questions such as on the DSTP because this assessment is structured to provide information on those students performing above proficiency.

Many comments were made regarding the need to be very clear on the terms used in the RFP. It was also discussed that we need to be broad enough in the language of the RFP so that we do not unnecessarily constrain what the vendor can do or suggest. Also, there was comment related to the ongoing costs to assessment. Any assessment that is developed has continuing costs for activities related to field testing, item development and infrastructure needs and changes.

Mrs. Woodruff stated that when the recommendations are made, there will be considerations stated as well.

Mrs. Woodruff provided a copy of a handout and Dr. Harter provided a schematic that outlined what was discussed at previous meeting. That guided the next part of the meeting.

Based on the conversation there was agreement on providing formative computer based adaptive assessments in reading and math to measure individual student growth over the course of the year in grades 2-8. The formative would be “on-demand”. Optimally, there would be Multiple Choice (MC) End of Grade (EOG) summative assessment in reading and math that could be given “on-demand” but at a certain time the student would be determined to have met or not met the expectations at that grade level. A summative on demand assessment would need to be very robust. An assessment with Student Constructed Response (SCR) questions and a writing assessment would be administered in the spring. This would provide the time needed to score those assessments. These scores would be combined with the MC EOG assessments and used for “high stakes” purposes. The formative would be low stakes and would not be used for any high stakes decisions (i.e. school or student accountability).

There was some agreement, but still outstanding questions related to the assessments in high school. There was tentative agreement on providing formative computer based adaptive assessments in reading and math to measure individual student growth over the course of the year in grades 9-12. The formative would be “on-demand”. The discussion then moved to the summative and End of Course (EOC) assessments. The task force members agreed that MC EOCs should be given in a minimum of the following content areas: writing, reading, math, science and social studies. Secretary Woodruff indicated we currently assess science and social studies in the 11th grade and based on the 11th grade content standards. There was also discussion related to block scheduling. It was discussed that the MC EOC assessment would be given at the conclusion of the course which could be mid-year. There was also discussion and agreement on having the Student Constructed Response and writing assessments in high school at grades where EOCs assessments were high stakes. It was decided that the writing assessment should be a stand-alone at the high school level. There was discussion that an EOC could count if taken in middle school as long as it was for purposes of providing credit for high school.

Mrs. Allen suggested that we need to be cognizant of the Graduation Requirement Committee’s intent to increase the rigor and provide more uniformity in our high schools and not select an assessment that measures standards below what we expect our students to know in order to graduate. There was a suggestion that the Graduation Requirement Committee provide input on defining what EOCs would be needed for graduation. There was discussion related to a recommendation of the Graduation Requirement Committee to require a foreign language credit. Mrs. Allen said this was being discussed as a second tier recommendation. Several members of the task force cautioned that we

need to be able to provide students with the supports to meet any new requirements and that the assessment system may gradually increase in rigor over several years. There was also discussion related to integrated course such as integrated math and integrated science that would require the EOC to look different perhaps than an Algebra II EOC. There was some agreement that whatever assessment is used as summative should measure at a minimum 10th grade standards in reading, writing, and math and grade 11th grade standards in science and social studies.

There was discussion related to the fact that some districts already have common assessments in certain course across the districts. The state assessment system does not preclude the districts from having their own common assessments in additional courses, but a comment was made that we should look to each other (districts and charter schools) to be more efficient when developing such assessments.

The discussion moved toward deciding if the EOCs should count toward the final grade. Vicki Cairns was going to check with the teachers to determine if they would have a concern with the EOC counting toward the final grade. However, in the task force there was preliminary agreement that the EOC should count toward the final grade, but the percentage (e.g. 5%, 15%) was not decided.

As it was getting late, Mrs. Woodruff said that we are making progress and that on May 17th we should be able to really work out some of the remaining details.

The next meetings are scheduled for:

May 17, 2006 at 8:30 a.m. – Townsend Building, DOE

May 25, 2006 at 1:30 p.m. – Townsend Building, DOE

HJR 4 with SA 2 – Assessment Task Force

Meeting Notes

May 17, 2006

Location: Cabinet Room, Delaware Department of Education

Time: 8:30 a.m.

Attendees: Bob Andrzejewski, Jean Allen, Vicki Cairns, Edie Corbin, Cindy DiPinto, Susan Haberstroh, Yvonne Johnson, Martha Manning, Wendy Roberts, David Sechler, Dorothy Shelton, David Sokola, Janine Sorbello, Robin Taylor, Nancy Wagner, Valerie Woodruff,

Public: Charlene Tucker, Elizabeth Reddin, Amy Drevna

Secretary Woodruff called the meeting to order. She distributed the meeting notes from the April 3rd, April 12th, and April 21st meetings. She also provided copies of the draft final report that had been provided by email to the task force members on May 16th. Mrs. Woodruff suggested that task force members review and make changes to the draft at this meeting. The group could then decide if another meeting was necessary or if email would suffice in finalizing the document.

The draft report was projected overhead and Nicole Quinn was to make changes during this meeting to expedite the process.

The group decided that the document needed a variety of changes. The group also decided to add a “Definitions” page to assist non-educators. One issue was the distinction between a “formative” and “summative” assessment. Robin Taylor gave a short description of the type of formative assessment for the purposes of this system. She commented that typically a formative assessment is used to inform instruction and that these assessments are not usually used for accountability purposes. They are generally given more than one time (usually two to three times), and they typically build toward something, such as the standards. For the purposes of this system, we are actually talking about a more formal formative assessment process. Summative assessments are typically a single event that shows how far the student has come from the beginning to the end of the year or period. These are more commonly used by states for accountability.

Vicki Cairns was concerned with the language of assessments as it relates to DPAS II. Secretary Woodruff commented that classroom assessments can be used as part of the DPAS II component related to student achievement.

Representative Nancy Wagner mentioned that the state needs to do more professional development on assessments. She commented that teachers do not see it as an assessment of themselves but rather as an assessment of the students.

A comment was made that the new mentoring program has a strong component related to assessment. This is only getting to a small population but the framework does exist.

Secretary Woodruff discussed the fact that Maryland has teachers score the assessments. In the Delaware request for proposal we can enter a variety of options for how the test should be scored. There were concerns voiced about the ability of districts to have teachers score the assessment Dorothy Shelton mentioned that the state already has

teachers doing benchmarking and it is a struggle to get people to do this. She is concerned about other things getting done, especially classroom time.

Nancy Wagner raised a concern that some legislators have commented that the scoring is currently done in Texas and think that it should be done in Delaware.

Robin Taylor stated that it could be priced both ways.

The discussion continued and moved onto the “description” portion of the document.

As drafted the report provides for two parts to statewide assessment. The constructed responses and writing would be in done in the March time frame with fewer days. The multiple-choice would be done as close to the end of the year as possible. Mrs. Woodruff commented that there had been consensus from previous meetings that the group did not want the entire assessment to be multiple-choice because it would not get the richness of knowing our students have critical thinking skills.

Dorothy Shelton wanted to make sure that the summary document included the issues of “less time” to get the scores and getting information in low performing students. It was decided that the summary document did address these issues.

Nancy Wagner stated that she would like the Department of Education to brief the legislators on the summary document. Yvonne Johnson stated that the group needed a shorter version of the document. The group decided an executive summary would be appropriate.

Nancy Wagner asked what “technology” meant in the document. Secretary Woodruff defined it to mean not only computer technology but the personnel costs and other supports needed. She stated that in order to get this done, a decent investment will need to be made. However, she said that technology is needed anyway whether it is for assessments or other purposes.

A discussion about end-of-course (EOC) assessments began. Select EOC assessments can be used for No Child Left Behind (NCLB) purposes while other EOC assessments can be used for other purposes. We need to use one that all students use, such as English Language Arts and Math. Robin Taylor mentioned to the group to not think grade-level but rather think subject. A student can take an end-of-course assessment at any grade-level and bank that score. Secretary Woodruff will have to fight for this with the United States Department of Education because NCLB requires students to be tested at the grade-level. In fact, Oregon and North Carolina are working on this right now.

Dave Sechler mentioned that the level of infrastructure discussed might have implications regarding the access for all students at the same time. Some schools might have a difficult time handling added infrastructure. Dorothy Shelton also mentioned that personnel are need for infrastructure. Therefore, infrastructure not only means technology, but also personnel.

Dorothy Shelton was concerned with the statement that implied some districts were not providing students with DSTP like assessments. She said, and Mrs. Woodruff concurred that some districts are doing a good job in developing assessments so that students are familiar with what is expected on the DSTP. Good assessments are those that are aligned and measure the standards. Mrs. Shelton also mentioned that she is uncomfortable with teachers scoring summative assessments. It is not a matter of trust, but rather an issue of priorities. Teachers should be in the classroom.

Nancy Wagner mentioned she wanted the group to think about the implementation of this new system and whether we have thought out and communicate any unintended consequences of this new assessment system.

Discussion followed related to the fact formative and computer-based assessments are different ways of assessing students. Mrs. Woodruff commented that with any change there is resistance and usually some discomfort for some. However, it was stated that it would be “less painful” if the schools have more infrastructure. In addition, there are security issues being investigated currently that happened during this last DSTP that will be a continuing concern as new technologies in general are created (i.e. text messaging, cell phone cameras).

Mrs. Woodruff commented that we have two priorities: formative assessments and upgraded changes to the summative assessments. The group decided that there must be a paper-and-pencil-based assessment even if computers are available. This could serve as an accommodation, plus may expedite changes to the summative portion of the assessment system. Paper and pencil assessments can be scanned and provide feedback in a timely manner.

Bob Andrzejewski. mentioned that it is important for teachers to see improvement as is done with the NWEA formative assessment currently being piloted in some of the Red Clay schools. This provides an incentive for the teachers when they are able to see growth in their classroom. Mrs. Johnson commented that as a parent, when students see that their scores improve it boosts their self-esteem.

Wendy Roberts mentioned that there will be transitional issues because the state will have to redo the cut scores. Robin Taylor mentioned that the state will still need to do reliability and validity studies.

The group decided that they wanted to review the edited document via email and then finalize it at the next meeting.

The next meeting is scheduled for:

May 25, 2006 at 1:30 p.m. – Townsend Building, DOE

Appendix B

Delaware and Federal School Accountability Requirements

	Delaware School Accountability	Federal No Child Left Behind (NCLB)
Academic Content Standards	Academic content standards in English language arts (reading, writing), mathematics, science and social studies in grade clusters K-3, 4-5, 6-8, 9-10/11 with grade level expectations at each grade, (excluding K and 1).	Academic content standards in English/language arts, mathematics, and science (2005-2006) at all grade levels 3 through 8 and one high school grade (can use grade level expectations)
Assessments	<p>Annual assessments DSTP reading, writing and math at grades 2 -10 (no writing grade 2); science and social studies at grades 4, 6, 8, and 11</p> <p>Delaware Alternate Portfolio Assessment (DAPA) for some students with disabilities</p>	<p>Annual assessments in reading/language arts and math in grades 3 through 8 and high school (2006). Until 2006, can assess one grade between 3-5, 6-9, 10-12. After 2006 must assess in each grade 3-8 and one high school grade between 10-12.</p> <p>Include science by 2008 at least once between 3-5, 6-9 and 10-12</p> <p>May measure other areas</p> <p>Alternate assessments must yield results in reading/language arts, math and science (2008); tie to state standards</p> <p>Include all students</p>
Academic Achievement Standards	5 levels of performance at grades 3-10 3 levels of performance at grade 2 (new for 2006)	At least 3 levels of performance – 2 at high levels (proficient and advanced) 1 at lower level (basic) for all grades 3 through 8 and high school
Reporting Assessment Information	<p>Individual student results at all grades, mailed to parents</p> <p>School, district and state results published annually in state summary (June) and school profiles (November) including disaggregated information by grade cluster, content area</p> <p>Release some test items when available – item samplers</p>	<p>Individual student results</p> <p>School, district and state results published annually including disaggregated information before the beginning of next school year</p> <p>Produce itemized score analyses</p>
Adequate Yearly Progress (AYP)	Same as Delaware's school accountability except that annual growth objectives are half of the school accountability targets	<p>Definition due Jan 31, 2003</p> <p>Determine starting point and annual growth objectives until 2014 as defined in law</p>

	Delaware School Accountability	Federal No Child Left Behind (NCLB)
	Students eligible to be tested – some exemptions	Must meet goal in each subgroup if there are a sufficient number of students in group Must have 95% participation (students tested compared to enrollment) in each subgroup (NO exemptions)
Student Accountability	Students with Performance Level 1 in grades 3 and 5 in reading; and 3, 5 and 8 in math to attend summer school unless other indicators shows proficiency to content standards. Students with Level 1 on summer school DSTP to be retained, students with Level II promoted with Individual Improvement Plan (IIP). Scores in reading, writing, math, science and social studies used in Diploma Index for distinguished diploma (2006 and 2007 only)	None
School Accountability	Every year based on ELA (reading 90%, 10% writing) and Math DSTP with other indicator. Other indicator for elementary and middle is increase in scale scores of lowest performing students in ELA and math; graduation rate for high school. State Progress determination based improvement on math, science, social studies, and reading assessments.	Every year; reading and math DSTP; one system; includes all students (95% participation) in a year
School Improvement	Two years of not meeting AYP in the same content area Next cycle school must meet all targets or be rated as under improvement Cannot get out of School Improvement unless meet AYP two years in a row	Two consecutive years of not meeting AYP Cannot get out of School Improvement unless meet AYP two years in a row Consequences can be delayed 1 year if meet AYP and under school improvement
Other Academic Indicators	Elementary and middle schools: improvement in reading and math for lowest performing students High school: graduation rate	Elementary and middle schools: state decision High school: graduation rate

Appendix C

Assessment Task Force Statement Draft

The **state assessment system** should **fairly and accurately communicate student achievement against standards** and **progress toward individual student goals** at **regular intervals**, provide **detailed information to inform instruction**, and provide educators with **information to tie student growth to strategic management decisions** that **improve student learning**. It should also provide information that will allow districts and the state to **measure the impact of instructional practices**.

It should **measure individual student progress** and have the necessary components to **measure teacher, student, and system performance**. It should **provide the public with a fair evaluation about the growth in student achievement occurring in each school**.

Appendix D

Delaware Student Testing Program (DSTP) Costs for the 2005-2006 School Year

The following pages provide a summary of the projected costs for the Delaware Student Testing Program (DSTP) for the 2005-2006 School Year. Please note that the contract year for the testing vendor bridges two DOE fiscal years. Therefore, some costs from 2005 and some from 2006 are reflected in these numbers. Additionally, many costs in the vendor contract are not isolated to an individual administration.

This report reflects the following:

- Costs listed are estimates based on overall proportions in various areas of the contract with the testing vendor (Harcourt Assessment, Inc.) plus Department of Education (DOE) costs for item development committees and other DOE support costs for the assessment program.
- Each administration includes estimates for the following four areas—
 - Item Development: Item Writing, Item Review, Bias and Content Review, Editorial Review, Composition, Printing/Manufacturing
 - Administration: Travel and Meetings, Materials Shipping, Contract Overhead, Consulting
 - Scoring: Scoring Development, Scoring Programming, Open Ended Scoring (Human Resources), Reader Training, Facilities/Computer Rental
 - Reporting: Reporting Programming, Quality Control, Printing, Materials, Shipping
- Some costs are testing vendor only, others are combined costs of DOE and the vendor – combined costs are footnoted.
- Item development costs include the DOE costs paid to Delaware educators who serve on the item writing and review committees.

Fall 2005 Administration of Grades 4 and 6 Science and Social Studies

Item Development	262,366 ¹	
Administration	93,365	
Scoring	389,019	
Reporting	77,804	
		<hr/>
		822,554

Spring 2006 Administration of Grades 2 – 10 Reading, Writing, Math

Item Development	1,909,562	¹
Writing Field Test (fall 2005)	877,163	²
Administration	555,741	
Scoring	2,778,706	
Reporting	444,593	
		<hr/>
		6,565,765

Spring 2006 Administration of Grades 8 and 11 Science & Social Studies

Item Development	262,366	¹
Administration	93,365	
Scoring	389,019	
Reporting	77,804	
		<hr/>
		822,554

Summer Administration of Grades 3, 5, and 8 Reading and Grade 8 Math

Item Development	14,820	
Administration	14,820	
Scoring	37,049	
Reporting	7,410	
		<hr/>
		74,099

¹This number includes Vendor and DOE costs.

²Writing field tests are conducted approximately every two years.

Retest for Diploma

Reading/Writing/Math	111,148	
Oct-05	37,049	
Mar-06	37,049	
Jul-06	37,049	
Science/Social Studies	111,148	
Oct-05	37,049	
May-06	37,049	
Jul-06	37,049	
		<hr/>
		222,296

Common Costs

Technology	130,000	
Technical Studies	155,387	
Bias Review	2,500	
Benchmarking	50,000	
Technical Advisory Committee	18,000	³
Item Samplers		
LEP Assessment	150,000	⁴
DE Alternate Portfolio Assessment	250,000	⁵
		<hr/>
		755,887

³DOE costs for TAC Stipends.

⁴Transferred to DOE School Improvement Work Group to offset costs of assessment of English Language Learners (ELL), including English language proficiency assessment and field tests of alternate assessments for English language arts.

⁵Transferred to DOE Exceptional Children and Early Childhood Work Group to partially offset costs of the Delaware Alternate Portfolio Assessment (DAPA)

Appendix E

Standards and Assessment Peer Review Guidance

April 28, 2004

Statutory and Regulatory Requirements for NCLB State Assessment Systems

Under NCLB, States must develop challenging academic standards that have the following characteristics:

- Be the same academic standards that the State applies to all public schools and public school students in the State;
- Include the same knowledge, skills, and levels of achievement expected of all students; and
- Include at least mathematics, reading/language arts, and, beginning in the 2005-2006 school year, science.

Academic **content** standards must specify what all students are expected to know and be able to do; contain coherent and rigorous content; and encourage the teaching of advanced skills. A State's academic content standards may either be grade-specific or may cover more than one grade if grade-level content expectations are provided for each of grades 3 through 8. At the high school level, the academic content standards must define the knowledge and skills that all high school students are expected to have in at least reading/language arts, mathematics, and, beginning in the 2005-06 school year, science, irrespective of course titles or years completed.

Academic **achievement** standards must be aligned with the State's academic content standards. For each content area, a State's academic achievement standards must include at least two levels of achievement (proficient and advanced) that reflect mastery of the material in the State's academic content standards, and a third level of achievement (basic) to provide information about the progress of lower-achieving students toward mastering the proficient and advanced levels of achievement.

For each achievement level, a State must provide descriptions of the competencies associated with that achievement level and must determine the assessment scores ("cut scores") that differentiate among the achievement levels. The State must also provide a description of the rationale and procedures used to determine each achievement level. Unlike content standards, which may address a cluster of grade levels, academic achievement standards must be developed for each grade and subject assessed, even if the State's academic content standards cover more than one grade.

With respect to academic achievement standards in science, a State must develop achievement levels and descriptions no later than the 2005-06 school year and must determine "cut scores" after the State has developed its science assessments, but no later than the 2007-08 school year.

Under NCLB, the State assessment system must have the following characteristics:

- Assessments must be aligned with State academic content and achievement standards, and they must provide coherent information about student attainment of State standards in at least mathematics and reading/language arts. Beginning in 2007-08, the system must also include assessments in science.
- The same assessment system must be used to measure the achievement of all students.
- The assessment system must be designed to be valid and accessible for use by the widest possible range of students, including students with disabilities and students with limited English proficiency (LEP).
- Initially, assessments must be administered annually to students in at least one grade in each of three grade ranges--grades 3 through 5, grades 6 through 9, and grades 10 through 12. Beginning in 2005-06, the mathematics and reading/language arts assessments must be given in each of grades 3 through 8 in addition to one of the grades 10 through 12.
- The assessment system must provide for--
 - Participation of all students in the grades being assessed;
 - Reasonable adaptations and appropriate accommodations for students with diverse learning needs, where such adaptations or accommodations are necessary to measure the achievement of those students relative to State standards; and
 - Inclusion of LEP students, who must be assessed in a valid and reliable manner and provided reasonable accommodations including, to the extent practicable, assessments in the language and form most likely to yield accurate and reliable information on what they know and can do in academic content areas, until such students have achieved English language proficiency; except that the reading/language arts achievement of any student who has attended school in the United States for three consecutive years must be tested in English.
- The assessment system must involve multiple approaches with up-to-date measures of student achievement, including measures that assess higher-order thinking skills and understanding of challenging content.
- Assessments must be valid and reliable for the purposes for which the assessment system is used and be consistent with relevant, nationally recognized professional and technical standards.
- The assessment system must be supported by evidence from test publishers or other relevant sources that the assessment system is of adequate technical quality for each purpose required under the Act.
- The assessment system must objectively measure academic achievement, knowledge, and skills without evaluating or assessing personal or family beliefs and attitudes, except that this provision does not preclude the use of constructed-response, short answer, or essay questions, or items that require a student to analyze a passage of text or to express opinions.
- Assessment results must be disaggregated within each school and district by gender, major racial and ethnic groups, English proficiency status, migrant status, students with disabilities as compared to students without disabilities, and economically disadvantaged students as compared to students who are not economically disadvantaged. Such disaggregation is not required when

the number of students in a category is insufficient to yield statistically reliable information or if the results would reveal personally identifiable information about an individual student.

- The assessment system must provide individual student interpretive, descriptive, and diagnostic reports that include individual scores or other information on the attainment of student achievement standards and help parents, teachers, and principals to understand and address the specific academic needs of students. These reports must be provided as soon as practicable after the assessment is given and in an understandable and uniform format.

Under NCLB, the statewide assessment system will be the primary means for determining whether schools and school districts are making adequate yearly progress (AYP) toward educating students to high standards. In determining the progress of schools, States must include scores of all students enrolled in the school for at least a full academic year. In determining the progress of school districts, States must include scores of all students enrolled in schools in the district for a full academic year, even if they have attended several different schools.

Because NCLB makes the State assessment system central to holding schools and districts accountable, this document focuses on the uses of the State assessment system at the school and district levels. Nevertheless, peer reviewers should note that the State assessment system is also required to report results at the level of individual students.

State Assessment System Design

A State may include in its academic assessment system either (or both) criterion-referenced assessments and assessments that yield national norms, provided that, if the State uses only assessments referenced against national norms at a particular grade, those assessments are augmented with additional items as necessary to measure accurately the depth and breadth of the State's student academic achievement standards.

A State that includes a combination of criterion and norm-referenced assessments in its assessment system must demonstrate that the system has a rational and coherent design that:

- Identifies the assessments to be used;
- Indicates the relative contribution of each assessment towards ensuring alignment with the State's academic content standards and toward determining the adequate yearly progress of each school and local educational agency (LEA); and
- Provides information regarding the progress of students relative to the State's academic standards.

A State's assessment system may employ either a uniform set of assessments statewide or a combination of State and local assessments. States using a combination of State and local tests must address issues of comparability and equivalency. For example, will proficiency on one local assessment be comparable to proficiency on another local assessment? Additionally, States must consider how they will aggregate to the State level the results from local assessments, as is required by NCLB.

States that choose to include a combination of State and local assessments will need to demonstrate that their system has a rational and coherent design that--

- Identifies the assessments to be used at the State and local levels;
- Indicates the relative contribution of each assessment toward ensuring alignment with the State's academic content standards and toward determining the adequate yearly progress of each school and LEA; and
- Provides information regarding the progress of students relative to the State's academic standards.

Further, a State that includes local assessments must also--

- Establish technical criteria to ensure that each local assessment addresses the depth and breadth of the State's academic standards; is valid, reliable, and of high technical quality; expresses student results in terms of the State's academic achievement standards; and is designed to provide a coherent system across grades and subjects.
- Demonstrate that all local assessments are equivalent in their content coverage, difficulty, and quality to one another and to State assessments; have comparable validity and reliability with respect to groups of students described in section 1111(b)(2)(C)(v); and provide unbiased, rational, and consistent determinations of the annual progress of schools and LEAs within the State.
- Review and approve each local assessment to ensure that it meets or exceeds the State's technical quality for assessments.
- Be able to aggregate, with confidence, data from local assessments to determine whether the State has made adequate yearly progress.

In implementing their assessment system, States have two main responsibilities: (1) they must develop, score, and report findings from State assessments, and (2) they must promulgate rules and procedures for local assessment systems if the State has such systems, as well as monitor them, to ensure technical quality and compliance with Title I requirements. The second function is particularly significant in assessment systems with strong local responsibility.

Section 1: A single statewide system of challenging academic content standards applied to all public schools and LEAs.

Reference in NCLB legislation: Sec. 1111(b)(1)
Reference in final regulations: Sec. 200.1

Overview

As the starting point for establishing a high quality assessment and accountability system under NCLB, States must develop a set of challenging academic content standards that define what all public school students in the State are expected to know and be able to do. A State's academic content standards are to be applied to all public elementary and secondary school students.

The table below provides a summary of the content, grade level, and timeline requirements for the academic content standards.

Content Area	Grade levels	Due	Notes
Reading/language arts	<ul style="list-style-type: none"> Each grade: 3 - 8; and Grade range: 10 - 12 	May 2003	<ul style="list-style-type: none"> If a State's standards cover grade ranges (e.g., 3 - 5 and 6 - 8) rather than the specific grades, 3 - 8, the State must develop grade-specific expectations in addition to its standards.
Mathematics	<ul style="list-style-type: none"> Each grade: 3-8; and Grade range: 10-12 		<ul style="list-style-type: none"> At the high school level, standards must define the knowledge and skills that are expected of all students prior to graduation. They may be linked to specific courses if all students must take these courses in order to graduate.
Science	<ul style="list-style-type: none"> Grade ranges: 3 - 5; 6 - 9; 10 - 12 	By the 2005 - 2006 school year	<ul style="list-style-type: none"> At the high school level, standards must define the knowledge and skills that are expected of all students prior to graduation. They may be linked to specific courses if all students must take these courses in order to graduate.

These standards must be rigorous and encourage the teaching of advanced skills. This means that a State should not adopt "minimum competency" standards or otherwise encourage low expectations for any students. Further, these standards must be coherent. That is, they must include only content that is meaningful with regard to the "domain", that is appropriate for the grade level specified, and that reflects clearly articulated progressions across grade levels.

Section 2: A single statewide system of challenging academic achievement standards applied to all public schools and LEAs.

Reference in NCLB legislation:	Sec. 1111(b)(1)
Reference in final regulations:	Sec. 200.1

Overview

To establish the level of achievement a State expects of all public schools and LEAs, the NCLB requires States to develop a set of challenging academic achievement standards for every grade and content area assessed. These standards are to be applied to all public schools and LEAs and ensure inclusion of those students with disabilities and students who are not yet proficient in English.

Achievement Levels

Academic achievement standards for each grade-and-content area combination must include at least three achievement levels, which the State may label 'proficient,' 'advanced,' and 'basic.' Of these levels, proficient and advanced must represent high achievement and basic must represent achievement that is not yet proficient. These labels may vary from State to State, such as "meeting and mastering" the State standards that would equate to the proficient and advanced labels as described in the statute. A State may use more than three levels, but must clearly indicate which level represents the proficient performance expected of all students.

Descriptors and Cut Scores

In addition to these levels, the State's academic achievement standards must include descriptions of the content-based competencies associated with each level. The State must also determine which specific scores on its assessments distinguish one level from another. These "cut scores" must be established through a process that involves both expert judgments and consideration of assessment results.

Alignment

As a set, the academic achievement standards must be aligned with the State's academic content standards in that they capture the full range and depth of knowledge and skills defined in the State's challenging, coherent, and rigorous academic content standards.

Timeline

Academic achievement standards in reading/language arts and mathematics for each of grades 3 through 8 and the 10-12 grade range must be in place by the 2005-06 school year. Academic achievement descriptors for science in grade spans 3-5, 6-9, and 10-12 must be in place by the 2005-06 school year and cut scores for science by the 2007-08 school year. States can develop the level and description components of the standards prior to the availability of assessment data that will be necessary to set the cut score components of these standards.

Alternate academic achievement standards

A State is permitted to define alternate achievement standards to evaluate the achievement of students with the most significant cognitive disabilities and to give equal weight to a limited number of “proficient” assessment results based on alternate achievement standards in calculating adequate yearly progress (AYP). Alternate achievement standards must be aligned with the State’s academic content standards (i.e., include knowledge and skills that link to grade-level expectations), must promote access to the general curriculum, and must reflect professional judgment of the highest learning standards possible for the group of students with the most significant cognitive disabilities. The State defines alternate achievement standards through a documented and validated standards-setting process similar to the process used to establish achievement standards on the regular assessments.

As a State expands the regular assessments to include grades 3 through 8, it must also provide alternate assessments at grades 3 through 8. If these alternate assessments are based on grade-level achievement standards, they will include the same grade-level content as the test for which they are an alternate. The assessment procedures may differ from the regular assessment (e.g., body of work or performance tasks instead of multiple choice) but proficiency on these alternates is comparable to proficient performance on the regular assessment for the same grade. The State must provide evidence of comparability and be able to aggregate the results with results from the regular assessment.

For alternate assessments in grades 3 through 8 based on alternate achievement standards, the assessment materials should show a clear link to the content standards for the grade in which the student is enrolled although the grade-level content may be reduced in complexity or modified to reflect pre-requisite skills. For each grade, the State may define one or more alternate achievement standards for proficiency.

For students with the most significant cognitive disabilities who are mainstreamed, the concept of alternate achievement standards related to grade level may be ambiguous. For practitioners, the question is whether the alternate achievement standards for this group of students must be clearly different from grade to grade. The alternate achievement standards should be defined in a way that supports individual growth because of their linkage to different content across grades. When examined across grades, however, the alternate achievement standards are not likely to show the same clearly defined advances in cognitive complexity as the achievement standards set for the regular test or an alternate assessment based on grade-level standards. States are expected to rely on the judgment of experienced special educators and administrators, higher education representatives, and parents of students with disabilities as they define alternate achievement standards and to define alternate achievement standards in a manner that provides an appropriate challenge for students with the most significant cognitive disabilities as they move through their schooling.

Section 3: A single statewide system of annual high-quality assessments

Reference in NCLB legislation:	Sec. 1111(b)(3)
Reference in final regulations:	Sec. 200.2, 200.3, 200.5

Overview

To ensure that States are able to evaluate whether all students are achieving to high levels, NCLB requires States to develop a single statewide system of high quality assessments. All public school students must participate in this assessment system, including those with disabilities and those who are not yet proficient in English, so States must make their assessment system fully accessible to all students, (see Principle 6 for more information about inclusion). States must employ the same assessment system for all their public elementary and secondary schools and students.

States must have the reading/language arts and mathematics components of their assessment systems in place by the 2005 - 2006 school year. These assessments must be administered annually to all students in each of grades 3 - 8 and at least once to students in the 10 - 12 grade range. By the 2007 - 2008 school year, States must also have in place their science assessments, which must be administered, annually, at least once in each of the 3 - 5, 6 - 9, and 10 - 12 grade spans. Assessments administered in the 10 -12 grade range in reading/language arts, mathematics, and science may be end-of-course tests so long as the associated courses, or combinations of courses, are ones that all students must take.

States must also have in place an alternate assessment designed for those students who are unable to participate meaningfully in the regular assessment, even with accommodations. For most students, results from this alternate assessment will be evaluated against the same grade-level achievement standards as the regular test for which it serves as an alternate. With respect to students with the most significant cognitive disabilities, however, the State may define alternate achievement standards. In addition, a few States have developed alternative assessments for use with LEP students. Results from these assessments must be judged against the same grade-level standards as the regular tests.

The assessments that make up the State's assessment system may either be criterion-referenced or an augmented form of a norm-referenced test. If the State uses only assessments referenced against national norms at a particular grade, those assessments must be augmented with additional items as necessary to measure accurately the depth and breadth of the State's academic content standards and express student results in terms of the State's student academic achievement standards.

A State's assessment system may include only statewide assessments, a combination of statewide and local assessments, or only local assessments.¹ However, if the State includes local assessments in its system, the State is responsible for ensuring that each of these assessments meets the rigorous criteria for technical quality and alignment specified in this document. The State must ensure that

¹ State law exception as stated in Section 200.4 of the July 5, 2002 Regulations and Section 1111(b)(5) of NCLB.

results from all local assessments can be aggregated meaningfully at the State level with one another and with scores from any statewide assessments.

In building its assessment system, a State must ensure that the information its assessments yield is coherent across grades and content areas. For example, information gained from the reading/language arts assessment at grade 3 should be clearly and appropriately relevant to information gained from the reading/language arts assessment at grade 4 and subsequent grades. This does not require use of tests that are vertically scaled, but does imply the articulation of the standards from grade to grade. The content of the assessments and the achievement standards should be articulated across grades.

Section 4: A system of assessments with high technical quality

Reference in NCLB legislation:	Sec. 1111(b)(3)
Reference in final regulations:	Sec. 200.2

Overview

The *Standards for Educational and Psychological Testing* (1999) delineates the characteristics of high-quality assessments and describes the processes that a State can employ to ensure that its assessments and use of results are appropriate, credible and technically defensible. The *Standards*, developed jointly by the American Psychological Association, the American Educational Research Association, and the National Council of Measurement in Education, has a history of 30 years of use by test developers and the courts.

Validity

As reflected in the *Standards*, the primary consideration in determining validity is whether the State has evidence that the assessment results can be interpreted in a manner consistent with their intended purpose(s).

The *Standards* speaks of four broad categories of evidence used to determine construct validity: (1) evidence based on test content, (2) evidence based on the assessment's relation to other variables, (3) evidence based on student response processes, and (4) evidence from internal structure.

1) *Using evidence based on test content (content validity)*. Content validity, that is, alignment of the standards and the assessment, is important but not sufficient. States must document not only the surface aspects of validity illustrated by a good content match, but also the more substantive aspects of validity that clarify the "real" meaning of a score.

2) *Using evidence of the assessment's relationship with other variables*. This means documenting the validity of an assessment by confirming its positive relationship with other assessments or evidence that is known or assumed to be valid. For example, if students who do well on the assessment in question also do well on some trusted assessment or rating, such as teachers' judgments, it might be said to be valid. It is also useful to gather evidence about what a test does *not* measure. For example, a test of mathematical reasoning should be more highly correlated with another math test, or perhaps with grades in math, than with a test of scientific reasoning or a reading comprehension test.

3) *Using evidence based on student response processes*. The best opportunity for detecting and eliminating sources of test invalidity occurs during the test development process. Items obviously need to be reviewed for ambiguity, irrelevant clues, and inaccuracy. More direct evidence bearing on the meaning of the scores can be gathered during the development process by asking students to "think-aloud" and describe the processes they "think" they are using as they struggle with the task. Many States now use this "assessment lab" approach to validating and refining assessment items and tasks.

4) *Using evidence based on internal structure.* A variety of statistical techniques have been developed to study the structure of a test. These are used to study both the validity and the reliability of an assessment. The well-known technique of item analysis used during test development is actually a measure of how well a given item correlates with the other items on the test. Newer technologies including generalizability analyses are variations on the theme of item similarity and homogeneity. A combination of several of these statistical techniques can help to ensure a balanced assessment, avoiding, on the one hand, the assessment of a narrow range of knowledge and skills but one that shows very high reliability, and on the other hand, the assessment of a very wide range of content and skills, triggering a decrease in the consistency of the results.

In validating an assessment, the State must also consider the consequences of its interpretation and use. Messick (1989) points out that these are different functions, and that the impact of an assessment can be traced either to an interpretation or to how it is used. Furthermore, as in all evaluative endeavors, States must attend not only to the intended effects, but also to unintended effects. The disproportional placement of certain categories of students in special education as a result of accountability considerations rather than appropriate diagnosis is an example of an unintended--and negative--consequence of what had been considered proper use of instruments that were considered valid.

Reliability

The term "reliability" is usually defined with synonyms such as consistency, stability, and accuracy. These terms all relate to the problem of uncertainty in making an inference about a score. As reflected in the *Standards for Educational and Psychological Testing*, the field now treats reliability as a study of the many sources of unwanted variation in assessment results. Those responsible for developing and operating State assessment systems are obliged to (1) make a reasonable effort to determine the types of error that may (unwittingly) distort interpretations of the findings, (2) estimate their magnitude, and (3) make every possible effort to alert the users to this lack of certainty.

The traditional methods of portraying the consistency of test results, including reliability coefficients and standard errors of measurement, should be augmented by techniques that more accurately and visibly portray the actual level of accuracy (Rogosa, 1995, Young and Yoon, 1999). Most of these methods focus on error in terms of the probability that a student with a given score, or pattern of scores, is properly classified at a given performance level, such as "proficient." For school-level or district-level results, the report should indicate the estimated amount of error associated with the percent of students classified at each performance level. For example, if a school reported that 47% of its students were proficient, the report might say that the reader could be confident at the 95% level that the school's true percent of students at the proficient level is between 33% and 61%. Furthermore, since the focus on results in a Title I context is on improvement over time, the report should also indicate the accuracy of the year-to-year changes in scores.

Other dimensions of technical quality

There are several other characteristics of State assessments that support valid interpretation and use of results.

Fairness/Accessibility The *Standards* identifies several sources of unfairness, including bias or unequal treatment of students in the assessment process or in the processes of reporting, interpretation, or use; and the lack of opportunity to learn to the standards. Unfairness most often appears at four points in the assessment process:

- The items or tasks do not provide an equal opportunity for all students to fully demonstrate their knowledge and skills.
- The assessments are not administered in ways that ensure fairness.
- The results are not reported in ways that ensure fairness.
- The results are not interpreted or used in ways that leads to equal treatment.

Comparability of results Many uses of State assessment results assume comparability of different types: comparability from year to year, from student to student, and from school to school. Although this is difficult to implement and to document, States have an obligation to show that they have made a reasonable effort to attain comparability, especially where locally selected assessments are part of the system.

Procedures for test administration, scoring, data analysis, and reporting Most States take great pains to ensure that the assessments are properly administered, that directions are followed, and that test security requirements are clearly specified and followed. Nevertheless, it is important they document the ways in which they ensure that their system does not omit any of these basics.

Interpretation and use of results Although this topic is closely related to that of validity, and is discussed in most of the other topics in this section, it is mentioned here because of its importance. Even if an assessment is carefully designed, constructed and implemented, it all can come to naught if users are not helped to draw the most appropriate interpretations and to use the results in the most valid ways.

Validation efforts continue throughout the life of the assessment. Evidence should continually be sought that the results truly reflect the goals of instruction, especially those related to higher-order thinking and understanding. Accurate data about the consequences of an assessment will, obviously, not be available until they have been implemented for a year or more. Research questions might ask: Are more students meeting the standards because the results led to the creation of a dynamic statewide after-school program? Are more students being retained in grade as a result of the assessment results? Are more teachers part of a long-term professional development program that improves the teaching of reading to low-achieving students?

Section 5: Alignment of Academic Content Standards, Academic Achievement Standards, and Assessments

Reference in NCLB legislation:	Sec. 1111(b)(1) and 1111(b)(3)
Reference in final regulations:	Sec. 200.2 and 200.3

Overview

A State's system of standards and assessments will provide useful information for valid accountability decisions and educational improvement only to the extent that all components of this system are aligned. If a State's assessments do not adequately measure the knowledge and skills specified in the State's academic content standards, or if they measure something other than what these standards specify, it will be difficult to determine whether students have achieved the intended knowledge and skills. As a result, it will be difficult to make appropriate policy, program, and instructional decisions meant to improve students' achievement. Further, if a State's assessments do not include items that cover the full range of the State's academic achievement standards, it may be difficult to determine whether students have reached the level of proficiency these standards describe.

Alignment encompasses several dimensions; demonstrating that an assessment system is aligned with a State's standards requires more than simply determining whether all the items on the assessment can be matched to one or more standards or whether each of the academic content standards can be matched to one or more items in the assessments. *Alignment is more than this two-way process.* To ensure that its standards and assessments are aligned, a State needs to consider whether the assessments--

- Cover the full range of content specified in the State's academic content standards, meaning that all of the standards are represented legitimately in the assessments; and
- Measure both the content (what students know) and the process (what students can do) aspects of the academic content standards; and
- Reflect the same degree and pattern of emphasis apparent in the academic content standards (e.g., if the academic content standards place a lot of emphasis on operations then so should the assessments); and
- Reflect the full range of cognitive complexity and level of difficulty of the concepts and processes described, and depth represented, in the State's academic content standards, meaning that the assessments are as demanding as the standards; and
- Yield results that represent all achievement levels specified in the State's academic achievement standards.

In addition to considering each of these aspects of alignment through a systematic development and review process, the State needs to also develop strategies for communicating to its education stakeholders how its standards and assessment are aligned. Parents, educators, and other

stakeholders need to know how assessment results are related to content-based expectations in order to understand and use test information effectively.

Each State must present evidence that its assessment system is aligned to its standards. Some alignment evidence is generated in the test development process, and documentation of the steps taken to ensure that items were drafted to reflect the full range of the State standards is appropriate verification of efforts to attain alignment. In addition, final alignment of assessments and standards following full implementation should be confirmed using one of several procedures (for example, review and comment by external subject-matter experts). Occasionally, documentation of alignment includes the process of re-verification if changes in tests were made to improve alignment.

In recent years, several methods of evaluating alignment between standards and assessments have been developed. A summary and comparison of alignment models can be found on the Council of Chief State Officers website at:

http://www.ccsso.org/Projects/alignment_analysis/models/418.cfm

When documenting the comprehensive aspects of alignment between standards and the State assessment system, the State should describe--

- The relationships between the structure of the standards and the structure of the assessments;
- The rationale for the overall alignment strategy, including a rationale for any standards either not assessed or not reported as part of the State assessment;
- The manner in which each standard is assessed, whether at the State, district, school, or classroom level;
- The manner in which alternate assessments based on alternate achievement standards are linked to the State content standards; and
- The type of information the State collects pertaining to each standard, and how the State monitors the quality of the assessment data collected at the local level, for all assessments that are part of the statewide system.

Section 6: Inclusion of all students in the assessment system

Reference in NCLB legislation:	Sec. 1111(b)(3)
Reference in final regulations:	Sec. 200.6

Overview

Just as its title indicates, one of the fundamental principles of the NCLB is the inclusion of all students in a state's system of standards, assessments, and accountability. By excluding any student or group of students from its assessment system, a state suggests that its high expectations apply only to some, but not all, students.

For some students with disabilities and for students who are not yet proficient in English, participation in the State's assessment system may require special considerations.² In all cases, however, decisions must be made regarding how an individual student will participate in the assessment system, not whether the student will participate.

To ensure that all students can participate fully in its assessment system, a State must allow:

- Participation in the regular assessment (limited English proficient students and students with disabilities); and
 - Participation in the regular assessment through the use of one or more approved accommodations (limited English proficient students and students with disabilities);
 - At least one alternate assessment, which may involve either or both of the following:
 - Participation in an alternate assessment that is aligned with the State's academic content standards and based on grade-level achievement standards (limited English proficient students and students with disabilities);
- and/or
- Participation in an alternate assessment that is based on alternate achievement standards (limited to students with the most significant cognitive disabilities).

Implementation of these options will require States to identify the needs of its special student populations so that it can appropriately address these needs. For example, for students who are visually- or hearing-impaired, the State needs to make available appropriate accommodations that will allow these students to demonstrate what they know and can do, as well as develop a system for ensuring that these accommodations are selected and used appropriately. For students with disabilities who cannot participate in the State's regular assessments, even with accommodations, the State must offer an alternate assessment that is based on the State's academic content standards, yields results in both reading/language arts and mathematics, and is designed and implemented in a manner that supports use of the results as an indicator of adequate yearly progress.

² Letter from Secretary Rod Paige dated February 20, 2004 on the flexibility in assessing new limited English proficient students and in measuring adequate yearly progress.

For students with limited English proficiency, the State must offer accommodations including, to the extent practicable, assessments designed to ensure that these students have an opportunity to demonstrate their academic knowledge and skills based on grade-level standards.

In addition to addressing the needs of students with disabilities and students with limited English proficiency, a State must take steps to ensure the participation of all migrant, otherwise mobile, and homeless students in its assessment system. This includes the accurate identification of migrant students and policies requiring assessment of all students, regardless of how long these students have been enrolled in the State.

It is important to note that as States continue to improve alignment between standards and assessments, the use of universal design principles holds great promise for designing and aligning standards, curriculum, instructional materials and strategies. Assessments that are designed to be valid and accessible for the widest possible range of students may help all students struggling to achieve, particularly students with cognitive disabilities, and would reduce the need for accommodations.

Section 7: An effective system of assessment reports

Reference in NCLB legislation:	Sec. 1111(b)(3)
Reference in final regulations:	Sec. 200.8

Overview

A State's assessment reports represent the culmination of all other aspects of its standards and assessment system. In these reports, a parent, educator, or other stakeholder should find answers to questions about how well a student or group of students is achieving, as well as important information on how to improve achievement in the future.

NCLB requires States to produce reports at the individual student, school, LEA, and State levels. At each of these levels, reports must include scores that are aligned with the State's academic content standards. Also, total test scores must be reported in relation to the performance levels defined in the State's academic achievement standards

Each of a State's reports should be produced and disseminated as soon as possible after each assessment administration. The individual student reports, at least, also need to be accompanied by interpretive guidance that will help parents and educators understand and be able to use the information the reports provide. States must ensure that this guidance is accessible to all parents.

States must carefully protect the data files containing student-level information that are produced following each assessment administration. When the State allows access to this information, it must do so in a way that maintains the confidentiality of each student's records.

Appendix F

What is the Evidence That Use of Multiple Choice State Tests Negatively Impacts the Teaching of Complex Reasoning and Higher Order Thinking Skills?

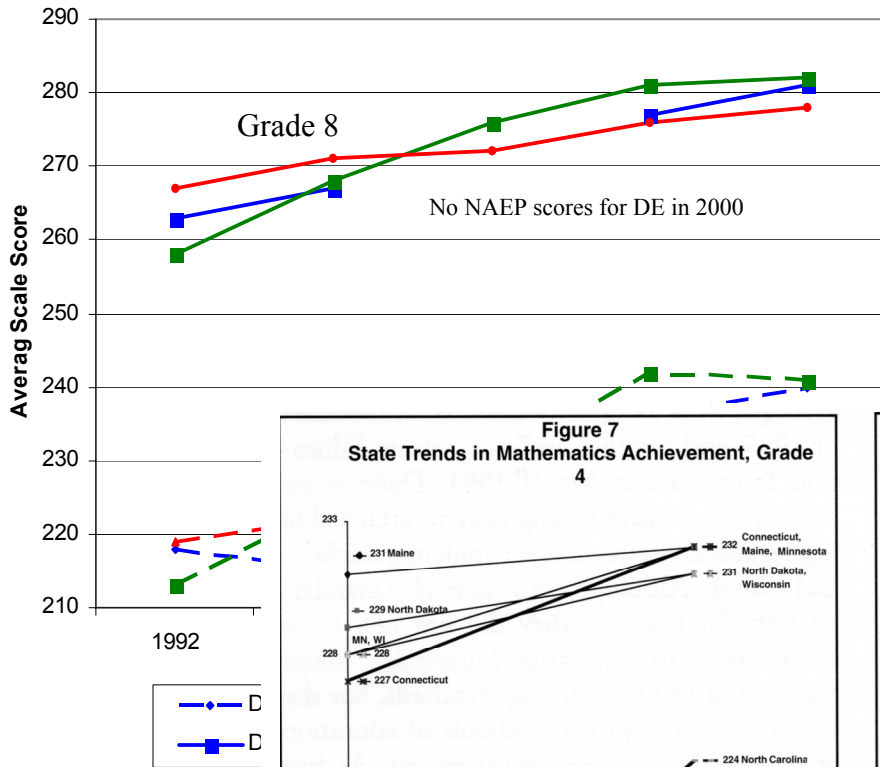
Background:

Many educators fear that use of solely multiple choice items on high-stakes tests will negatively impact instruction such that student performance on more complex tasks that require high order thinking will decline. Student-constructed response items and essays, some assert, must be included on the highly-stakes tests to ensure the teaching and learning of such skills. What is the evidence to support this?

One means of exploring this question is to look at performance over time on the NAEP, which is widely recognized as a very high quality and rigorous assessment that requires the use of higher order thinking skills. By comparing average statewide performance on the NAEP between Delaware, which has always included student-constructed response items and an essay, with North Carolina, which is demographically very similar but has always used solely multiple choice items, we can test this assumption.

Demographics of Delaware and North Carolina Public School Enrollment		
	Delaware	North Carolina
Student Characteristics		
Percent eligible for free/reduced lunch:	33.8%	44.5%
With Individualized Education Programs (IEP):	14.6%	14.2%
Percent in limited-English proficiency programs:	3.4%	4.5%
Racial/Ethnic Background		
White:	57.3%	58.3%
Black:	31.9%	31.6%
Hispanic:	7.9%	6.7%
Asian/Pacific Islander:	2.6%	2.0%
American Indian/Alaskan Native:	0.3%	1.5%
School/District Characteristics		
Pupil/teacher ratio:	15.2	15.1

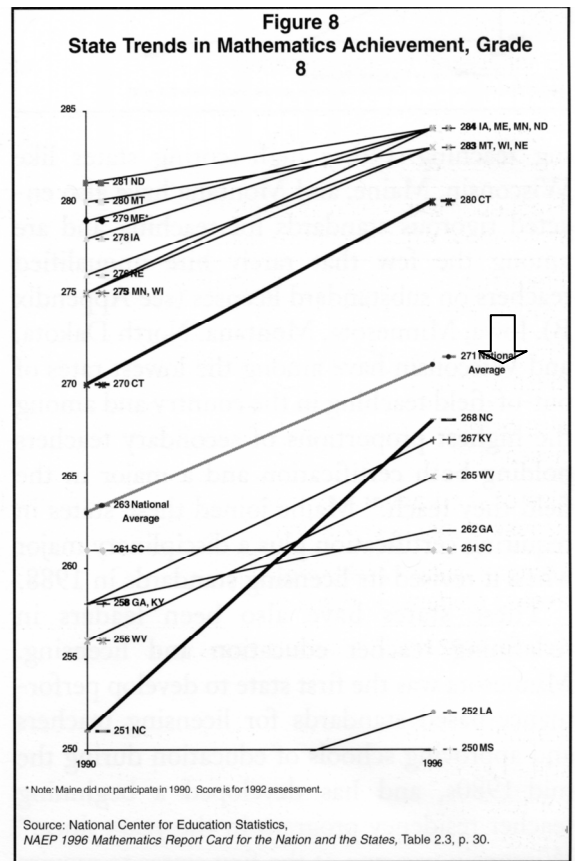
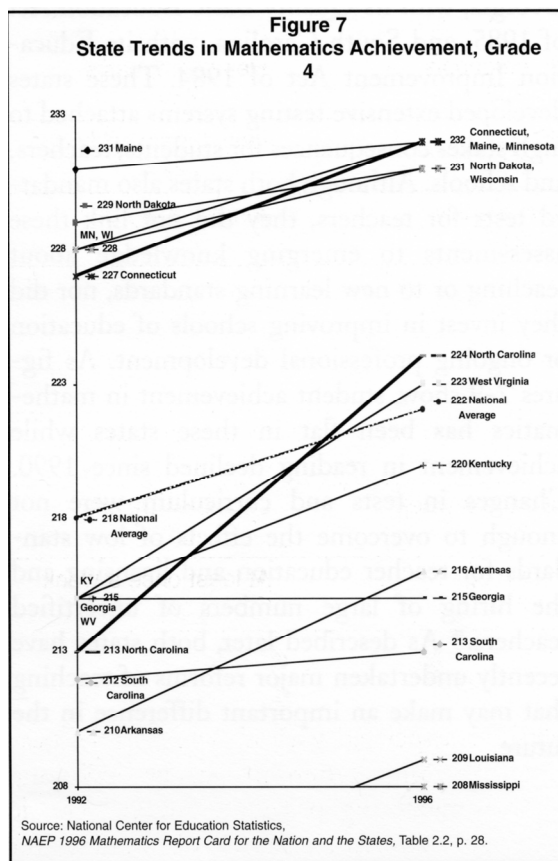
Improvement in NAEP Scores: Mathematics



Might it Make Sense to Redirect Dollars from High- Cost Assessment Items into Improving Teacher Quality?

In 1997, the National Commission on Teaching and America's Future released a report entitled "Doing What Matters Most: Investing in Quality Teaching." The report identified the states making the most rapid gains on

NAEP, and then investigated what those states had been doing that might explain those gains. They concluded that the high rates of improvement in student learning in both North Carolina and Connecticut were attributable to the investments in efforts to improve teacher quality, including teacher preparation, mentoring, and professional development.




DOING WHAT MATTERS MOST: INVESTING IN QUALITY TEACHING

As the previous page shows, North Carolina has continued to make NAEP gains equal to or better than Delaware's, despite use of low-cost multiple choice tests. How do Delaware's efforts to improve teacher quality compare to North Carolina's? Probably the most comprehensive look at this is done annually by Education Week in their annual Quality Counts report.

	1997*	1998*	1999	2000	2001	2002	2003	2004	2005
DE	C	C	C+	C-	D+	D+	D+	C	C
NC	C	C+	A	B	B+	B+	B	B	B

* In 1997 and 1998, Quality Counts rated "Quality of Teaching." Beginning in 1999, they rated, "Efforts to improve teacher quality."



Delaware Statewide Student Assessment System

Design for the Next Generation

Delaware Task Force on the Future State Assessment System

Ellen Forte, Ph.D.
April 3, 2006



Elements of Purpose

...a system of assessments that fairly and accurately measure student achievement against state content standards and that is designed to support the improvement of student learning.

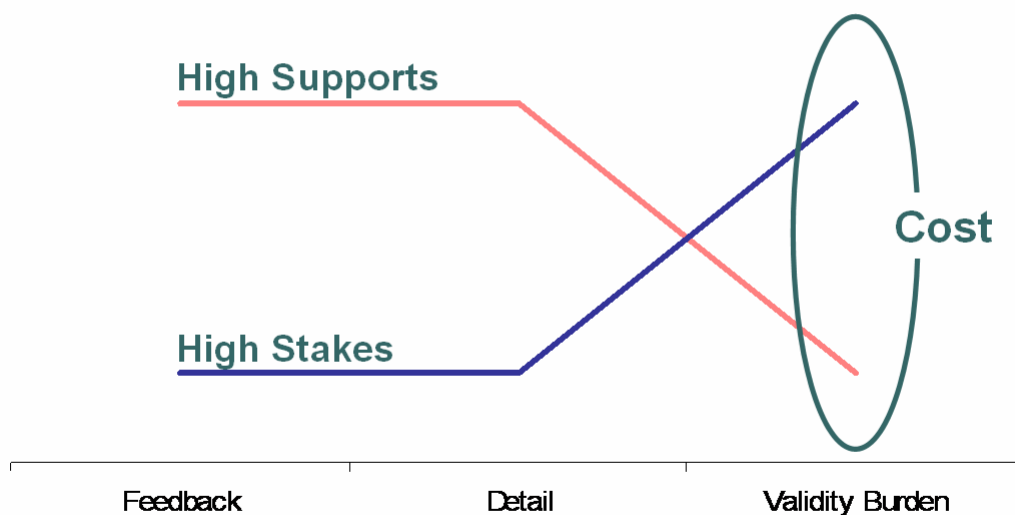
This system must provide useful information for use in:

- Instructional decisions
 - Curricular/program decisions
 - Accountability decisions for—
 - Students
 - Schools and Districts
 - State
- Support**
- Stakes**

From Purpose to Design: *Supports and Stakes*

- Supports and stakes are linked so the assessments should be, too
- Supportive information
 - frequent and relatively immediate feedback
 - detailed enough to support specific instructional actions
 - validity burden is lower because results are combined with other information
- Stakes information
 - feedback provided annually or in relation to a valued event or period
 - overall statement of performance in relation to expectations
 - validity burden is very high (and higher for student stakes than for school or district stakes)

From Purpose to Design: *Supports and Stakes*





Validity Burden

- Alignment with standards
- Opportunity to learn/remediate
- Accessibility for all students
- Secure administration and maintenance of items and forms
- Reliable and accurate scoring



From Purpose to Design: *Supports and Stakes*

- Multiple assessments or multiple components are necessary to fulfill all purposes
- Instruments that are meant to provide supportive information should be administered at regular intervals or on demand, provide detailed feedback, and be accompanied by extensive support for students, educators, and parents
- Instruments that are meant to provide stakes information should be administered (or yield results) once annually at the same time for all students, provide overall feedback, and be accompanied by extensive support for schools and districts

● ● ● | Elements of Purpose

...a system of assessments that fairly and accurately measure student achievement against state content standards and that is designed to support the improvement of student learning.

This system must provide useful information for use in:

- Instructional decisions ← **District**
- Curricular/program decisions ← **District**
- Accountability decisions for—
 - Students ← **State**
 - Schools and Districts ← **State**
 - State ← **State**

● ● ● | From Purpose to Design: *Shared responsibility*

- | | |
|---|--|
| ○ The state components of the system are used to make accountability decisions. | ○ The district components of the system are used to support student instruction. |
| ○ The state has the higher validity burden. | ○ Districts have the higher professional and curricular development burdens. |

From Purpose to Design: *Features*

Shared

- Alignment with standards
- Overall item specifications and bank
- Computerized delivery options
- Scoring criteria
- Immediate feedback on select portions
- Interactive reports and data files
- Responsibility for promotion/retention and graduation decisions

Not shared

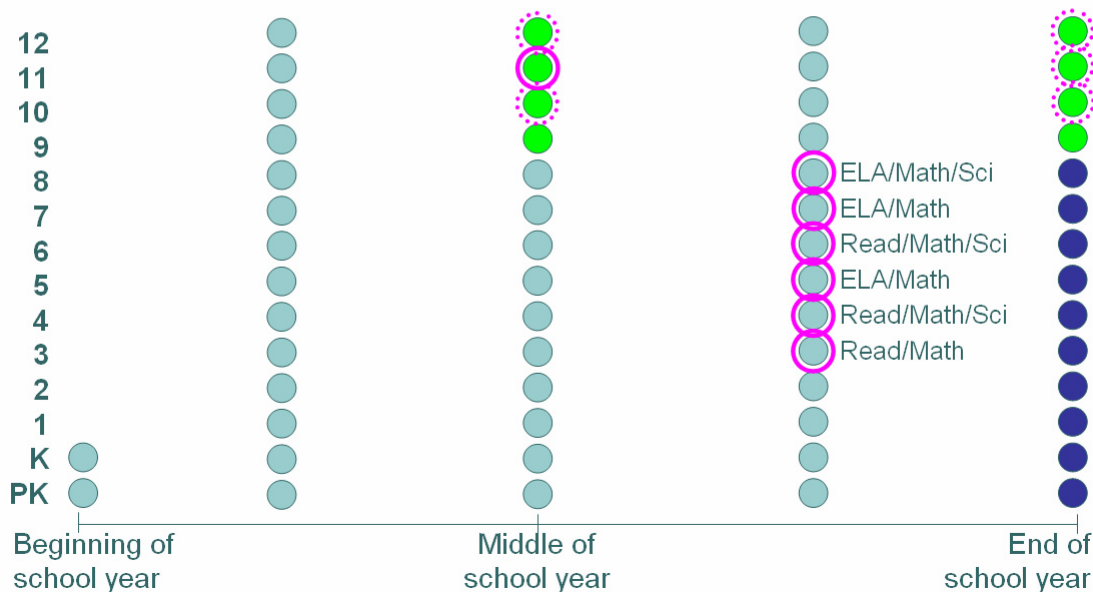
- Alignment with curriculum
- Items used for high stakes tests
- Computer-adaptive delivery option
- Local scoring
- Total scale scores and achievement levels (MC & CR)
- Responsibility for using results to support instructional and curricular decisions and actions
- Responsibility for grading and staff-related decisions



One option



- = district-administered formative assessment
- = district-administered end-of-grade assessment
- = district-administered end-of-course assessment
- = state-administered accountability assessment
- = pre/retest



Appendix H

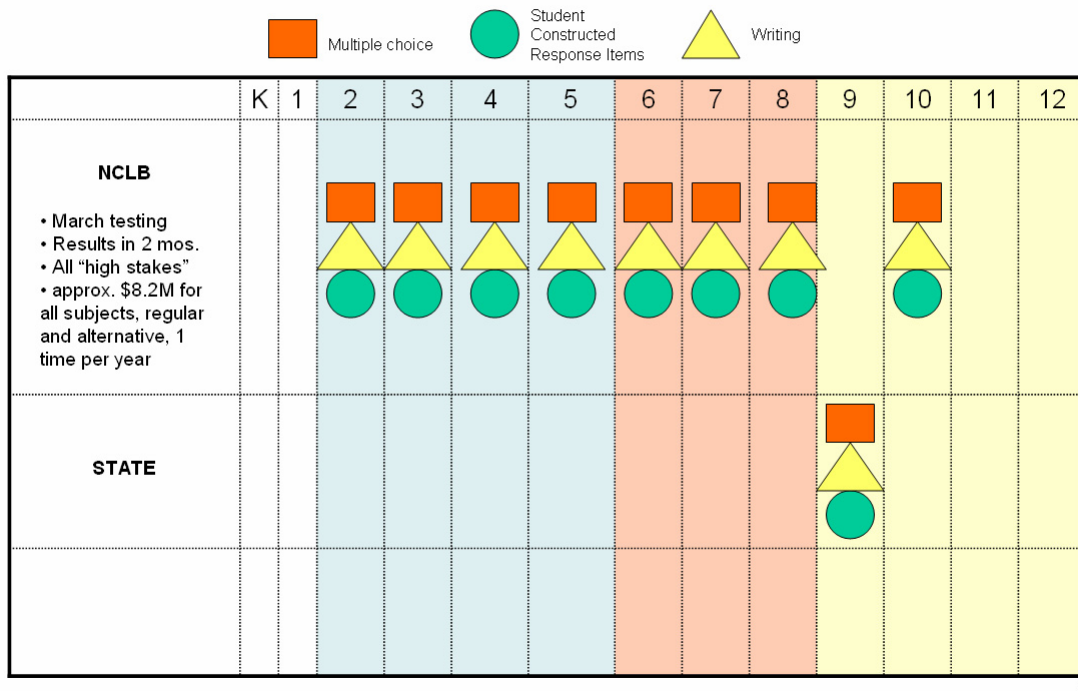
“Straw Man” Proposal for the Next-Generation State Assessment System

Submitted March 31, 2006
to the Delaware Task Force on the Future State Assessment System
created by HJR 4

by task force members:

Nancy Doorey – Delaware School Boards Association
Bruce Harter – Delaware Chief School Officers
Vicky Cairns, Delaware State Education Association
Martha Manning, Delaware Charter Schools Network
Edie Corbin, Metropolitan Wilmington Urban League

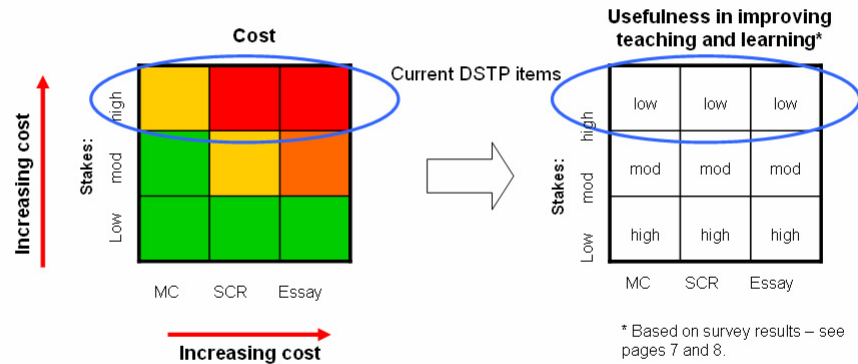
The Current State Assessment System: Built for Accountability Purposes



2

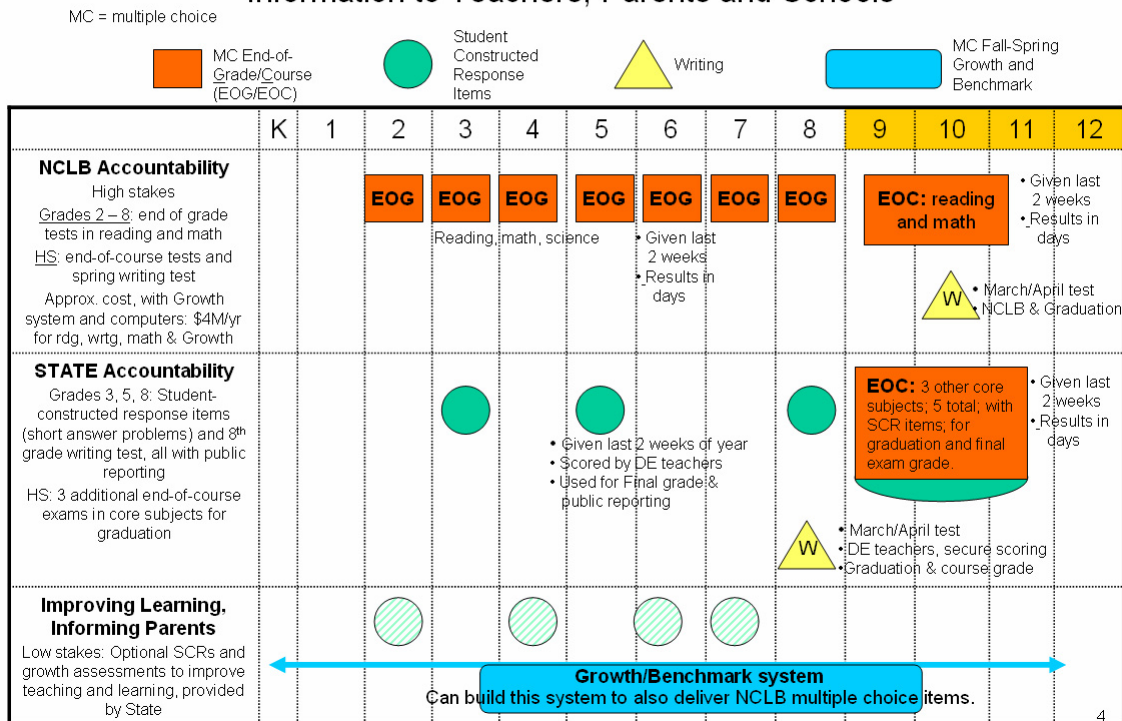
Current State Assessment System

All 3 items types are used at each grade, 2 – 10, and all are high-stakes, which increases cost and reduces value to teachers and parents.



3

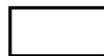
“Straw Man” Proposal: Reduce NCLB Costs, Test Full School Year, and Provide Better Information to Teachers, Parents and Schools



The Assessment Calendar

G/B 45 – 60 minute test on computer

Grade	Sept	Oct	Nov	Dec/Jan	Feb	Mar	Apr	May/June
2	G/B			G/B				EOG G/B
3	G/B			G/B				EOG G/B
4	G/B			G/B				EOG G/B
5	G/B			G/B				EOG G/B
6	G/B			G/B				EOG G/B
7	G/B			G/B				EOG G/B
8	G/B			G/B		W		EOG G/B
9-12	G/B			G/B		W		EOG G/B



Box indicates that the 2 tests are delivered as a single 2-phase computer-based test for rapid scoring and rapid diagnostic feedback to teachers and administrators.

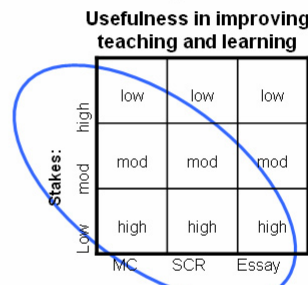
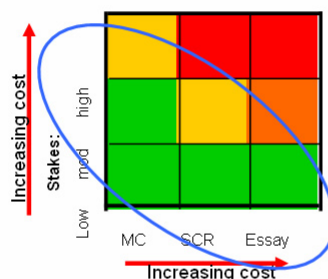


5

Written Explanation

Reduce the assessments tied to NCLB in order to move tests to end of year and reduce costs. (This model has been used by North Carolina since the 1980's and has been approved by US DOE.)

- Grades 2 – 8: multiple choice end-of-grade tests given during final 2 weeks of year, results within days
- High School: for NCLB use only a reading/ELA and a math end-of-course exam given during final 2 weeks, and count scores toward final exams, as well. Students can take tests multiple times during grades 9 – 12.
- Additional benefits: reduced staff time at schools for testing preparation, reduced testing time



Keep Student-Constructed Response items for DE accountability at former grades of 3, 5, 8 and high school. Score within state using DE teachers (powerful professional development) and a secure scoring process. Report results to public.

DE Graduation Requirement: Passing scores on 5 end-of-course exams in core subjects, including writing. Tie student-constructed response items to final grade only, to keep cost and scoring time down. At least 1 EOC exam from student's concentration/career pathway.

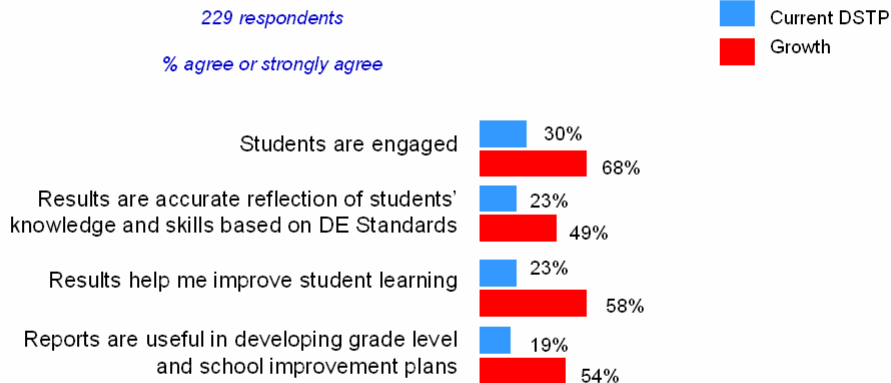
Reallocate savings to assessment tools that improve teaching, learning and parent information:

- Fund a customized statewide student growth and benchmark system that provides much more detailed and frequent student learning information to teachers and parents, and has higher educator confidence. Use this platform to deliver NCLB portion (2-phase computer based test) to eliminate printing costs and speed up scoring.
- Embed within the recommended curriculum some high quality end-of-unit mini-assessments, including student constructed response items and anchor papers. (Smithsonian Science a good example)

6

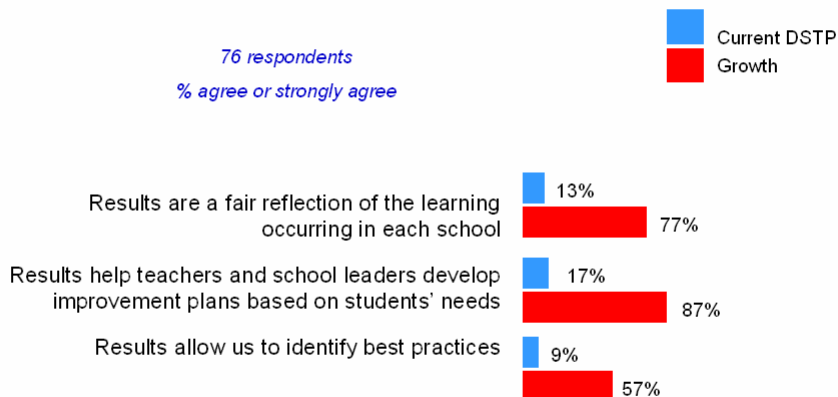
An informal and very early survey of Delaware educators who had been using the growth assessment system for 7 months

Teachers' Views



7

Administrators' and Instructional Coaches' Views



8

Strong Support for Keeping Student Constructed Response Items

% "important" or "very important"

	Teachers	Admins
For NCLB accountability	11%	22%
For DE accountability, but not NCLB	17%	24%
Required, but scored within districts	53%	45%
Not at all	19%	10%



9

Desired Attributes of Next State Assessment System

% "important" or "very important"

	Teachers	Admins	In Current	In Proposed
Detailed instructional feedback for each student	89%	95%	no	yes
Fall-spring measure of individual student growth	87%	97%	no	yes
Results within days	83%	98%	no	yes
Benchmark assessments that predict end-of-year performance	65%	93%	no	yes
Computer-adaptive tests to increase score accuracy	80%	94%	no	yes
Electronic reports: grade level, classroom, subject, year		100%	no	yes



10

An Evaluation of the Alignment of Three Formative Assessments to Delaware Grade Level Expectations

Reading and
Mathematics,
Grades 3, 5, 8, and 10

Phoebe C. Winter

April 21, 2006

Tests Reviewed

- Six test publishers invited to submit formative assessments for review
- Three test publishers submitted assessments:
 - Northwest Evaluation Association
 - Measured Progress
 - Harcourt Assessment, Inc.

NWEA: Measures of Academic Progress – Delaware Version

- Typically administered as a CAT and currently used in some DE school districts
- Multiple-choice items only
- NWEA created forms for the alignment review using the following criteria:
 - the most frequently exposed items in DE at each grade level
 - match to DE GLEs in content and balance
 - test length of 40 items for reading, 50 for math

MP: Progress Toward Standards

- Developed to address core content across states
- Multiple-choice and constructed response items
- Reading – 42 m-c items, 2 c-r items
- Math – 48 m-c items, 2 c-r items

Harcourt: Stanford Learning First – Class Views

- Custom-developed for each state – no form available for DE
- Grades 3, 5, and 8 only
- Reading – FL test
 - Grades 3 and 5: 36 m-c items;
 - Grade 8: 30 m-c items, 2 c-r items
- Math – IL test
 - Grades 3, 5, and 8: 39 m-c items, 1 c-r item

Webb's Alignment Review Procedure

- Review is conducted by educators with expertise in the content area
- Four characteristics of alignment
 1. Categorical concurrence
 2. Depth of knowledge consistency
 3. Range of knowledge correspondence
 4. Balance of representation

Criteria for Evaluating Alignment

- **Categorical concurrence**
 - At least 6 items per standard
- **Depth of knowledge consistency**
 - At least 50% at or above the DOK of the GLEs
- **Range of knowledge correspondence**
 - At least 50% of GLEs measured by at least one item
- **Balance of representation**
 - Index value of at least .70

Sample Results

Table 4. Summary of Alignment Results for Grade 3 Reading by Test

Test	Standard	Categorical Concurrence	Depth-of-Knowledge Consistency	Range of Knowledge	Balance of Representation
MAP	2	YES	YES	NO	WEAK
	4	NO	WEAK	NO	YES

Results – Reading

- Across tests and grade levels
 - the items matched one or more GLE
 - the tests met the categorical concurrence and depth of knowledge criteria for Standard 2
- No test met the categorical concurrence criterion for Standard 4 in Grade 5
- Progress Toward Standards met the categorical concurrence criterion for Standard 4 in Grades 3, 8, and 10

Results – Reading

- Progress Towards Standards, Grades 3, 8, and 10, are sufficiently aligned to DE GLEs
- Progress Towards Standards (G5), Measures of Academic Progress – Delaware Version (all grades), and Stanford Learning First – Class Views (all grades), have major weaknesses in their alignment to DE GLEs

Results – Mathematics

- Across tests and grade levels
 - a few items did not match a GLE
 - in most cases, the tests did not meet the categorical concurrence criterion for Standards 5 through 8 (process standards)
- All tests had major weaknesses in their alignment to DE GLEs



Appendix J

Next-Generation State Assessment System

DRAFT

G/B Growth tests: computer adaptive, 45 – 60 minutes

EOG End of Grade/Course test: computer-based, 45 – 90 minutes



Constructed response items – 2 per content area



Writing assessment

Grade	Sept	Oct	Nov	Dec/Jan	Feb	Mar	Apr	May/June
2 – 8	G/B	Available on demand up to three times			EOG G/B	per year plus end of summer school		EOG G/B
						CR W		
9-12	G/B					CR W		EOC
	Optional, available in reading and math, up to 3 times per year							

Grades 2-8: Growth assessment (computer adaptive) in reading and math given on demand at beginning of school year and up to two more times, plus end of summer school
 EOG tests in reading and math given on demand up to three times per year, with Growth assessment coupled to; EOGs can be given through last week of school, results within 2 days
 Grades 3, 5 and 8: Science and Social studies EOG tests given in final month of school
 Constructed response items: 2 short response items per content area, plus writing assessment, given in spring

Grades 9 – 12 Growth assessments available for use in reading and math
 At least 5 EOCs required for graduation: courses TBD, but at least 1 writing, 1 math, 1 reading, 1 science, 1 social studies
 Taken last 2 weeks of course, with 2nd version available in final week for those who do not pass on first try
 Results within 2 days
 Constructed response items: 2 short response items per content area, plus writing assessment, given in spring

DRAFT

Appendix K

Assessment System Components

Two major components

- Locally administered formative assessments
- State annual summative assessments

Reading, Writing and Mathematics Annually in grades 2 thru 8

- At least 3 times per school year (beginning, middle and end)
- Multiple choice
- Formative and adaptive to provide instructional guidance for individual students
- Extended response items earlier than end of year to be used in conjunction with “final scores” on formative multiple choice items or state summative

Science and Social Studies

- Formative annually embedded as part of science and social studies curriculum (Coalition Science kits – Statewide Recommended Curriculum) -
- Summative in grades 3, 5 and 8

End of course (EOC) assessments in high school

- As close to end of year as possible
- Multiple choice
- Extended response items earlier than end of year to be used in conjunction with end of year multiple choice portion (writing assessment)
- Content area?
- Grade levels?

Scoring:

- Use technology to extent possible
- Use Delaware teacher for extended response scoring to extent possible
- Release some extended response items annually to be used for professional development and for parent and student information

Reporting:

- Formative: provide reports for teachers, students and parents each time assessment is given
- Extended response: release some items for use by teachers, for students and parents
- End of course: provide reports for teachers, students and parents

Use of each component:

- Informing instruction
- Focusing professional development
- Accountability (which components?)

